1992

The effects of training on understanding the law of large numbers.

Jill Hendrickson Lohmeier
University of Massachusetts Amherst

THE EFFECTS OF TRAINING ON UNDERSTANDING THE LAW OF LARGE NUMBERS

A Thesis Presented
by
JILL H. LOHMEIER

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

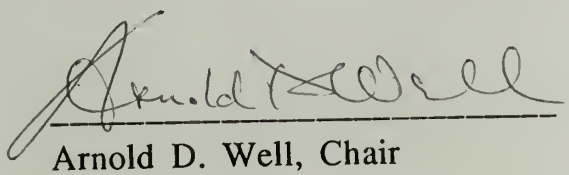
February 1992

Department of Psychology

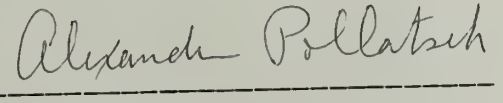
THE EFFECTS OF TRAINING ON UNDERSTANDING
THE LAW OF LARGE NUMBERS

A Thesis Presented
by
JILL H. LOHMEIER

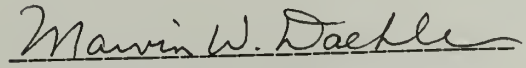
Approved as to style and content by:



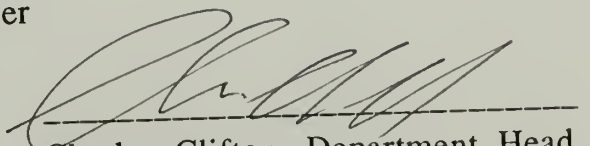
Arnold D. Well, Chair



Alexander Pollatsek, Member



Marvin W. Daehler, Member



Charles Clifton, Department Head
Department of Psychology

ACKNOWLEDGEMENTS

I would like to thank the members of my committee, Marv Daehler, Cliff Konold, and Sandy Pollatsek for their time and their shared wisdom. I would especially like to thank my committee chair, Arnie Well for his undying patience, availability and advice.

I would also like to thank my friends and family for all of their support and encouragement. In particular, my parents deserve many thanks for the encouragement they have continually given me, as well as the confidence they have shown in my abilities.

Finally, I would like to thank Steve for more things than I can name. Most importantly, thank you for being a patient, interested and supportive friend.

TABLE OF CONTENTS

LIST OF TABLES

Table

Page

CHAPTER I

INTRODUCTION

Throughout the problem solving and reasoning literatures many claims have been made about the types of rules people use to solve everyday problems. One major question asked is whether people use abstract domain-independent rules or concrete domain-specific empirical rules in reasoning. Many scholars, beginning with Plato, have believed that people rely on abstract rules in reasoning and problem solving in many different domains. This theory, known as the formal discipline, proposes that people can learn abstract rules through education in subjects such as Latin, logic, and mathematics and apply them to concrete everyday domains.

Piaget has been an influential modern proponent of this formalist view, suggesting that throughout cognitive development abstract rules and schemas are acquired which may be used to solve general problems in the future. One major difference, however, is that Piaget believed that abstract thinking is acquired through cognitive development, not through direct education. However, since Piaget's theory was proposed, others have found that children do not progress through stages in the orderly manner he suggested. They do not always perform similarly on tasks that represent the same stage of cognitive development. This suggests that reasoning does not occur solely on an abstract level, rather

children's reasoning is based on specific rules that they have learned.

Others, beginning primarily with Thorndike, (1906; Thorndike & Woodworth 1901) have also suggested that people do not possess such abstract rules. On the contrary, people possess rules that are used only in the specific domains in which they were learned. These rules are used to deal with concrete events. Thorndike's research led him to believe that the transfer of learning was entirely dependent on the occurrence of identical elements in the training and testing situations.

More recently, Wason's work (1966) has also been important in supporting this empirical notion. Wason's selection task requires subjects to determine what information is necessary to confirm or reject a rule. The rule his subjects were given was "If a card has an A on one side, then it has a 4 on the other." The subjects were then shown four cards on which was written either an A, a B, a 4 or a 7. They were then asked to turn over the two cards which would confirm this rule. The correct answer would require subjects to turn over the A and the 7. Wason found that people made many errors when attempting to reason abstractly about these arbitrary symbols and tended to give the answers A and 4. However, in subsequent research he found that when people were given similar problems in more familiar and concrete domains so that the task made sense based on the content of the problem, they were able to solve selection task problems with little trouble (Evans 1982; Johnson-Laird, Legrenzi & Sonino-Legrenzi 1972; Wason & Johnson-Laird 1972). Such research has

led many to support the empirical view of reasoning. That is, while people may form rules, the rules they form are not abstract rules that are easily applied across domains. Rather, they are domain-specific rules based on specific experiences (Griggs & Cox 1982; Reich & Ruth 1982).

Not only are the types of rules people use to solve everyday problems of interest, it is also important to ascertain whether or not people easily abstract and use these rules. Pedagogically, it is important to know whether these rules may be taught, and how this is best done. Subsequently it is important to learn how readily the rules are then used in various domains.

Strong transfer effects have been difficult to demonstrate in the problem-solving literature (Gick & Holyoak 1980, 1983; Singley & Anderson 1989). For example, Gick and Holyoak (1980, 1983) gave subjects a story to read about a military problem in which a general wants to capture an enemy fortress. The fortress, located at the center of several incoming roads, is difficult to attack because any large group on any road leading into the fortress will cause mines to explode. The general is therefore required to send his troops in small groups up each road. After reading this problem and answer, the subjects were then asked to solve Duncker's radiation problem (1945). In this problem a doctor needs to perform surgery on a tumor through the use of rays that can destroy human tissue if they are of sufficient intensity. The doctor can not send in enough rays to destroy the tumor at one place or else other tissues will be destroyed. Only 20% more subjects were able to solve this problem after receiving the

military example than those subjects who received no analogous example.

Cheng and Holyoak (1985) reported similar results about the effects of training on transfer of logical rules. They found that an entire course in formal logic did students little good in avoiding errors when they were asked to solve conditional selection problems such as Wason's selection task. However, when subjects were given more familiar concrete problems such as, "If a person is drinking alcohol, then he/she must be over 21", they were able to identify the correct options for checking this rule when given the following four options: someone drinking a soft drink, someone drinking alcohol, someone over 21, someone under 21. Cheng, Holyoak, Nisbett and Oliver (1986) reasoned that the formal logic training was ineffective because in reasoning people use practical permission schemas, not formal rules. These pragmatic schemas are generalized sets of rules which are defined in relation to classes of goals and the relationships to those goals. These schemas therefore allow people to reason correctly about the situations for which they hold these schemas, but do not allow them to transfer this ability. This theory is somewhat different than both the empirical and formalist viewpoints. It suggests that although people do not use abstract rules, they also do not merely possess concrete domain-specific rules. Rather they possess pragmatic structures which can be used in several situations.

The effects of training have been investigated in several areas other than logical reasoning. One important area is that of statistical reasoning. Tversky and Kahneman conducted research

investigating how people reason statistically in everyday situations (1971, 1974; Kahneman & Tversky 1972). They found that instead of using normative statistical rules, subjects used heuristics such as "representativeness", "availability", and "anchoring and adjustment". Additionally, they also first noted people's disregard for sample size when solving probabilistic problems. They concluded that through the use of the representativeness heuristic people make probabilistic judgments with no regard for sample size.

Understanding the effects of sample size is extremely important in understanding statistics, thus, there have been a number of investigations of how training affects peoples' understanding of the law of large numbers. The law of large numbers (LLN) states that as a sample's size increases, the statistics of the sample become less variable and therefore become a better estimate of the corresponding population's parameters.

Fong and Nisbett (Fong, Krantz & Nisbett, 1986; Fong & Nisbett 1991) have conducted several studies dealing with the training of the LLN, which they have argued support the formalist theory of reasoning. Fong et al. (1986) concluded that training had positive effects on students' abilities to solve problems in which the LLN is applicable. In several studies Fong and Nisbett (Fong et al. 1986; Fong & Nisbett 1991) found that brief training sessions on the LLN led to significant improvement rates on future LLN problems.

In the 1986 study, Fong et al. tested four types of training. These included a <u>demand</u> condition, an <u>abstract training</u> condition,

a guided induction condition, and a full training condition. In the "demand" condition subjects were given a one-sentence definition of the LLN and one example problem. The abstract training condition involved a brief written explanation of the LLN followed by a verbal presentation. This presentation included explaining the LLN by drawing various samples from a population of red and blue gumballs in an urn. The guided induction consisted of a paragraph that introduced the LLN, followed by several example problems. Following the problems were explanations of how to use the LLN in those problems. The full training condition consisted of both the abstract and guided induction training methods. These four training groups were compared to a control group that received no training.

Following the training, students were asked to solve several problems that represented various structures and domains within the LLN. The problems were divided into three content domains. The first, "probabilistic", consisted of problems in which an element of randomness was made obvious. For example, a random generating device was included, or the sample was said to be random. The "objective" problems required subjects to make judgments about objective data. For these problems, data were generally included about two different situations, and subjects were asked to make a judgment based on that data. The last domain, "subjective", consisted of problems that contained subjective data about which the subjects were asked to make judgments. For example, one problem described a situation in which a student needed to decide to attend one of two schools,

based on a large amount of information from friends, or his own small sample of first impressions of the schools.

Each domain included six different problem structures. In the first, subjects were required to draw conclusions about a population from one small sample. The second structure required a comparison between a small sample and a large sample. In the third, subjects were required to explain why an outcome with extreme deviation was not maintained in later samples. Structure four is said to be like structure two, the difference being that the large sample was drawn from a population related, but not identical to the target population. With the fifth structure, subjects were required to compare a large sample to a theory that was not actually founded in the data. That is, the theory sounded reasonable, but no evidence was presented to support it. The last structure was for the false alarm problems. In these, conclusions were made from a large, but highly biased sample. An answer was considered to be a false alarm if the subject indicated that a larger sample was needed to make a conclusion. Thus, these problems were aimed to detect any overgeneralization of the LLN. The examples in the guided induction training were from the objective problems domain.

Based on a three-point coding system, in which an answer was given a score of 1 if it was deterministic, a 2 if it was a poor statistical answer and a 3 if it was a good statistical answer, Fong et al. (1986) found that students who received the complete training were able to produce 22% more statistical answers (scores of 2 and 3) than control groups averaged across the above problem

domains (42% vs. 64%). The percentage of statistical answers of high quality (the proportion of scores of 3 out of all of the statistical answers, 2s and 3s) was also higher after training (control 54% vs. full training 70%). Students with either the rule training or example training alone also did better than the control group. Their responses included 56% and 54% statistical answers, respectively, with 67% and 66% respectively of these answers considered to be of high quality. They concluded that these results supported the formalist theory of reasoning because subjects were able to use an abstract concept more frequently after having received either abstract training or examples training alone.

In their 1991 study, Fong and Nisbett were primarily interested in determining whether subjects were using domain-independent rules to solve problems or were solving problems by analogy from example problems. In their first experiment, they trained subjects on the LLN, using example problems from either the domain of sports or of ability testing. They then tested their subjects with problems from both domains either immediately or after a two-week delay. They found that when subjects were tested immediately, there was no interaction effect between training and testing domains. However, a significant interaction between training and testing domains was found after the two-week delay. In two of the experiments subjects were given a questionnaire, which probed subjects' memories for the problems and training they had received in the first testing session, at the end of testing sessions. Because no subjects were able to recall many details about any of the problems they received during the

first testing session, Fong and Nisbett interpreted the delayed testing interaction as indicating that subjects' memories for the abstract rules were sparked by familiar problems,

Fong and Nisbett (1991; Fong et al. 1986) have relied on the formalist point of view to explain their results. They argue that people do in fact have abstract inferential rules which they use to solve everyday problems, and that abstract training manipulates these rules. The main results they provide to support this theory are as follows:

1) The abstract rule training alone led to improvements in both the frequency and quality of statistical answers (from 42 to 56 percent and 54 to 67 percent).

2) Abstract rule training was effective across problem structure domains.

3) Examples training (guided induction) generalized across domains (probabilistic and subjective) that were not included in training (Nisbett, Fong, Lehman & Cheng, 1987).

Thus, they conclude that with training in one domain the problems in several domains became easier for subjects to solve correctly. Before commenting on the conclusions found in the Fong et al. papers, it is important to note other research that has been conducted in this area.

Well, Pollatsek and Boyce (1990) have also conducted training studies on the LLN. One type of training used in their study involved computer supplemented sessions in which sampling distributions were explained to subjects. The subjects were then led to create their own sampling distributions. Like

Fong et al. (1986), Well et al. (1990) also found that training had significant effects on the subjects' abilities to solve problems. However, Well et al. (1990) also found that after training, subjects still did not understand the effects of sample size on the variability of the sample average. After training subjects were asked to predict how increasing the sample sizes in sampling distributions from 10 to 100 would affect the variability in the sampling distribution of the means. Subjects could not correctly predict that larger samples would decrease the variability. Thus, they concluded that although students' performance on problems dealing with the LLN improved with training on sampling distributions, the students still did not fully understand the LLN.

Some of the differences found between the Fong et al. (1986) and Fong and Nisbett (1991) conclusions and those of other studies, such as the Well et al. (1990) research, lead to many questions for teachers of statistics and scientists studying the learning of statistical concepts. Some of the important issues about the above mentioned studies that should be examined carefully, and which will be discussed here are: the coding systems used to evaluate performance; the testing materials; the important results; and the conclusions that should be drawn from the results.

The coding system in the Fong et al. work (1986; Fong & Nisbett 1991) has repeatedly been based on a three-point scoring system. According to this system, the subjects received one point for giving an answer that made no use of any statistical concept, that is, an entirely deterministic response. A 2 was given for "poor statistical responses". That is, if subjects mentioned any statistical

idea, even if it was wrong, they received two points. This score was also given to those who mentioned the LLN (which was in the instructions) but did not give an explicit answer. A 3 was given for a "good statistical response". Correct use of a statistical response yielded this score. Fong and Nisbett (1991) state that training not only led to a greater quantity of statistical answers (scores of 2 and 3), it also led to higher quality statistical answers (that is, more 3s). This argument does not follow from the data. The control group's mean was 1.5. Immediately after training, the mean score for test questions from the same training domain was approximately 1.85; 1.90 for sports, and 1.83 for ability testing problems. This indicates that most students did not, in fact, receive mostly twos and threes. As Ploger and Wilson (1991) have recently pointed out, these numbers indicate that the majority of students, regardless of training condition, are not providing good statistical reasons. In fact the scores obtained by Fong and Nisbett could have occurred if students had simply been repeating something like, "The Law of Large Numbers applies here" in their answers. Ploger and Wilson (1991) also suggested that this could be true, pointing out that while a quarter or less of the subjects in the Fong and Nisbett (1991) study correctly applied the LLN, more than three quarters were able to recall the law well enough to receive the highest score possible in the quality coding system.

One interesting difference between another recent Well et al. (1992) LLN training study and the similar Fong et al. (1986) study is that the training group in the Well et al. study did an average of 52% better in immediate testing and 57% better in delayed testing

than the control group. This is much greater than the 22% frequency difference in the Fong et al. 1986 study. If with a larger improvement rate Well et al. found that the students still did not have an adequate understanding of the LLN, it does not seem possible that the trained subjects in the Fong et al. study (1986) could have had an adequate understanding. Another important difference is that the control group in the Fong et al. study (1986) correctly answered 42% of the questions, with 52% giving "quality" answers. In Well et al. (1990) only 23% of the control group gave correct answers. It is important to look at the differences in testing materials given such varying results.

Beyond the differences in results, it is important to note the differences in training. The Fong et al. (1986) and Fong and Nisbett (1991) abstract rule training included several one-sentence descriptions of the LLN in terms such as, "the larger the sample, the better it is in estimating the population", "the larger the sample is that you draw, the better that sample is in estimating the population", etc. The Well et al. (1990) training included some instruction on how sample size affects variability and also attempted to explain how statistics taken from samples produce sampling distributions. The increased variability of the distributions of sample means when the samples are small as compared to when they are large was also addressed directly. Although the more complex training may have only confused subjects about the issue of variability, it must have at the very least made them aware of the issue. The Fong et al. (1986) and Fong and Nisbett (1991) instructions mentioned bigger deviations

with small samples, but did not address the issue of variability in much detail. With such training one may not understand the importance of variability in the LLN. The Fong et al. (1986) subjects were able to reproduce the fact that the larger samples are able to better estimate the population than small samples are. This is an important fact, but being able to reproduce it is not an indication of having an accurate abstract rule for the LLN.

An adequate understanding of the LLN needs to include several concepts. These concepts include understanding the differences between population, sample, and sampling distributions, especially in terms of the effects of different sample sizes on sampling distributions. Additionally, students should understand why big samples can be used to make estimates about population parameters, and why they are better than small samples in terms of the variability of the samples' statistics. Another important idea that students should have at least a basic understanding of is regression toward the mean. Finally, in order to have an adequate understanding of the LLN, it is essential to understand why random samples are necessary, in terms of efficiency and accuracy.

An important issue to consider in describing adequate learning of the LLN, is the distinction made by the Gestalt psychologists between senseless and meaningful learning (Katona, 1940). They referred to senseless learning as the kind studied by the associationists, that is, merely learning by memorizing. Meaningful learning occurs through a deep understanding of the underlying structure in problems. They claimed that meaningful

learning would encourage transfer much more frequently than senseless learning. Perhaps the difference in the results of the Fong et al. (1986) and Well et al. (1990) studies lies in this issue. That is, while in the Fong et al. (1986) study subjects showed improvement with training, senseless learning may have been all that occurred. While it is doubtful that the subjects in the Well et al. (1990) study truly had a deep understanding of the underlying structure of the LLN problems, perhaps these subjects experienced the beginning stages of more meaningful learning.

In continuing research Well et al. (1992) have also tested understanding the LLN by using problems in which the LLN is not applicable, but with a limited amount of understanding, might appear to be. These so called "lure" problems were intended to serve the same kind of purpose as the Fong et al. (1986; Fong & Nisbett 1991) false alarms. An example of a lure problem, the fish hatchery problem follows:

"The manager of a fish hatchery monitors data about the length and weight of trout that are raised in the hatchery tanks. This information is important because it has been found that if the trout are too small when they are released into rivers and lakes, the survival rate will be low. Since there are thousands of trout in the tanks, the data are obtained by taking random samples of the trout in each tank. From past data, they know the average weight of trout at a certain age is one pound. On a given day, two employees each take a random sample of trout at that age to measure. However, one takes a sample of 10 and the other takes a sample of 50. They weigh each trout in the sample and compute the percent of trout that are less than 3/4 of a pound.

Which would be true?

(1) The percent of trout less than 3/4 pound should be greater in the small sample than in the large sample.

(2) The percent of trout less than 3/4 pound should be greater in the large sample than in the small sample.

(3) There is no reason to expect that the percent of trout less than 3/4 pound would be greater in one sample than in the other.

Please write down the reasons that you selected the answer you did. "

Superficially, the above problem looks like a sampling distribution problem; however it is, in fact, asking about the distribution of scores within samples. Therefore, a subject who was applying a "bigger is better" heuristic without much understanding might select answer (1), instead of the correct answer (3). Well et al. (1992) found that subjects who received some training were much more likely than subjects that received no training to make errors on lure problems. (On the above problem, the control group made 0% errors, the trained group made 92% errors.) Most no-training subjects were able to answer these lures correctly, but were not able to answer the test questions correctly. This supports neither an empirical nor a formalist view. The former view suggests students would learn a domain-specific rule which was not transferrable across domains. The latter suggests students would learn the abstract rule and correctly apply it only where it was applicable. However, what occurred in this study is that the subjects abstracted part of the rule and then attempted to apply it in most of the problems.

Perhaps when people do form abstract rules, they simply do not know how to apply them. Kahneman and Tversky have distinguished between errors of comprehension and errors of application (1982). Errors of comprehension occur because people do not correctly understand a rule. Errors of application occur when people know a rule, but do not know how to correctly apply it to certain problems. Although it is possible the students originally made errors of comprehension, after training they seem to have also made errors of application. Perhaps students do not simply fall into one of these two categories. Rather, they may make errors of both types because they possess only a partial understanding of the concept.

Several possibilities have been mentioned about how people understand the LLN and other similar statistical concepts. The LLN is a concept that everyone encounters at some point in their lives in practical everyday situations, whether they are aware of it or not (Nisbett, Krantz, Jepson & Kunda 1983). Some everyday examples include trying to understand how the information gained from a poll estimates information about a population. Other examples may arise through participation in sports. This can come through experience with outliers or regression to the mean situations, such as a champion team losing, a home run hitter striking out, a poor batter getting a hit, or a specific player's average regressing to the mean. Apart from sports, the concept may be encountered in many other ways. For instance, gardeners realize that a very small sample from one's garden may not be indicative of the quality of all of the plants in that garden, and

certainly not indicative of the quality of all of the gardens in a town that year. Students may realize that if they only ask three classmates how they did on an exam, they may not get an accurate indication of the class's average or distribution of overall performance. Therefore, as Fong et al. (1986) and Fong and Nisbett (1991) have suggested, subjects may in fact have a general concept of the LLN that can be made more clear through training. It is also possible that while people do have some understanding of the concept, they do not have an in-depth understanding that they easily apply across domains. Based on what was found in the Well et al. (1990, 1992) studies, it can be seen that an in-depth understanding is not something that is quickly taught in brief training sessions, and then accurately applied across domains.

Many have suggested that there are several aspects involved in the concept of the LLN. (e.g., Fong & Nisbett 1991; Holland, Holyoak, Nisbett & Thagard 1986; Kunda & Nisbett 1986; Well et al. 1990). Although it is simple to teach people that in drawing samples "bigger is better", it is much more difficult to teach them what this means in terms of variability, predicting from populations to samples, standard deviations, etc. Fong et al. (1986) employed several structures which touched on different approaches to presenting LLN problems; however, the problems representing all of the structures, other than the false alarms, could correctly be answered with a superficial understanding of the LLN. The false alarm problems used in the study did not actually test the subjects' abilities to decipher when the law is applicable and when it is not. Rather, they test one's ability to

acknowledge biased information. An example of a false alarm problem, the brewery problem follows:

"A brewery buys nearly all of its reusable glass bottles from a local glass manufacturer. One summer, however, the local company is unable to deliver enough bottles, and the brewery orders a shipment from a large glass manufacturer that distributes its products nationwide. On the first day that these new bottles are used, however, the bottle-filling machinery has to be stopped four times because of jamming, and as a result, production for the day is unusually low. (Ordinarily the brewery does not experience more than one jamming stoppage per day and frequently there are none at all.) The foreman is worried about the new bottles. He decides to test the new bottles produced by the national manufacturer carefully. He randomly selects 300 cases of these new bottles and instructs the bottle-filler operators to record carefully each jamming incident. Meanwhile, company mechanics carefully lubricate and check adjustments on the bottle-filling machinery. When they are finished, the bottle-filling machinery is running more smoothly than it has in years. During the next 2 days, the 300 cases of new bottles are fed to the machine. There are only two jamming incidents, one each day. The foreman concludes that there is in fact little or no real disadvantage of the new bottles with respect to jamming of the bottle-filling machine.

What do you think about the foreman's reasoning? Give reasons for your answers."

Understanding that biased information is not good information is an important aspect of understanding the LLN, however, it is not exactly a difficult or counterintuitive notion. People will quickly assert that biased information is not worthwhile. If subjects had been given slightly more stringent false alarm problems in their study, their false alarm error rates might have been higher. They do not provide a test for overgeneralization on topics closely related to the LLN, as the lure problems in the Well et al. studies did (1992). It seems rather feasible that some subjects may have

slipped through the false alarms. That is, they may have answered these questions as they would have before training. They could have gotten these questions correct without understanding why. Therefore, I do not think that the Fong et al. (1986) and Fong and Nisbett (1991) studies adequately tested for understanding of the LLN. Thus, subjects may in fact have an abstract rule which they apply across domains, but the rule may not be correct or may be only partially adequate. Subjects may attempt to apply this rule across domains when the problems are not obviously unrelated to the rule.

I am also not convinced that Fong et al. (1986) and Fong and Nisbett (1991) have provided strong evidence that training is naturally and easily applied across domains. The instructions in their studies told the subjects that the LLN may be helpful in solving some of the problems given to them. Subjects with little understanding of the LLN could easily try to recall a basic statement from the training, and with a simple "bigger is better" do quite well on all of the problems. The Fong et al. (1986) training materials for the abstract rule condition contain seven statements defining the LLN (See Appendix A). All of these statements are variations of the statement "the larger the sample, the better it is in estimating the population". Using those materials is a good way to be sure that the students realize the importance of the LLN and remember a simple definition of it. However, remembering this simplistic definition is not necessarily indicative of actually understanding of the law.

Fong and Nisbett (1991) concluded that after a two week delay, subjects did not perform as well in their untrained domains as in their trained domains because their memory for the rule they learned in training was sparked by problems similar to those they received in training. Fong and Nisbett also stated that their results were not due to subjects' memories for specific examples. Fong and Nisbett (1991) support this conclusion by suggesting that the students did not know the two sessions were related. It is not implausible that the thought did not occur to the subjects. When the subjects returned they were not given further training; however, as before, the questions they were given included the instructions, "In many of the problems you may find the law of large numbers is helpful." Upon reading that, subjects may have grasped for any idea they could remember, such as the simple rule that big samples are better than small ones. This is enough of an understanding to receive at least two of the three possible points in the Fong and Nisbett (1991) scoring system. Although this can be considered an abstract rule, it is only one component of the LLN. The fact that subjects may be able to remember it does not warrant saying that they retained a good understanding of the LLN.

Thus, while previous research on peoples' understanding of the LLN has been important and informative, it has left us with many questions still to answer. Some of the important questions which still need to be studied include the following: Can statistical concepts such as the LLN be quickly taught in the form of accurate abstract rules that are easily and appropriately applied across

domains? Do people easily learn heuristics and overapply them without fully understanding what the statistical concepts mean? What aspects of the LLN are being taught in brief training sessions and are easily learned? What is missing? Is there a way in which subjects can be taught some things in a brief session that would lead them to use the LLN when appropriate, but also to understand that sometimes it is not appropriate and why? It is important for researchers interested in learning, especially in the learning of statistical concepts, to address these questions in their research. In the present paper some of these issues have been addressed through studying the effects of several types of training on students' understandings of the LLN.

CHAPTER II

EXPERIMENT

Purpose

The present study was conducted in order to establish a more adequate test of what subjects learn from various types of training than had been previously provided. In order to establish whether training was as successful as was previously claimed by Fong and Nisbett(1991, Fong et al., 1986), the present study includes some training conditions similar to those used in the Fong and Nisbett research, along with test questions from these and other studies. A training method similar to one used in Well et al. research (1990) was also tested in order to ascertain whether subjects who received this type of training were better able to understand different aspects of the LLN than were subjects who received another type of training, such as that from the Fong and Nisbett research.

The present study attempts to discover what subjects understand about the LLN beyond simply being able to recall it. This includes attempting to look at what subjects understand about the role of accuracy and variability in the LLN after brief training sessions by investigating the kinds of rationales which subjects give for their answers. Additionally, developing a more adequate coding system than that used by Fong and Nisbett was thought to

be imperative in order to ascertain whether subjects can adequately apply the LLN. As Ploger and Wilson (1991) have suggested, it seems that although the Fong and Nisbett subjects (1991) were better able to remember the LLN after training, they were still not able to consistently apply the rule. Therefore, the intentions for the present study were to provide questions that would encourage subjects to directly deal with various statistical concepts related to the LLN, such as accuracy or variability, and to determine if they would consistently apply the LLN.

The purpose was to see if students can transfer a learned rule across domains and apply it to several types of everyday statistical problems, as the formalist theory would suggest. In order to give subjects the chance to do this on their own, the instructions did not indicate that the LLN would be helpful in answering the problems.

My prediction was that in this training study the formalist theory would not be strongly supported. I expected that when subjects were given LLN problems to solve after training, they would attempt to transfer some knowledge about the LLN to various types of problems. However, they would attempt to transfer a basic rule, such as bigger is better, without fully understanding the LLN. Additionally, they would attempt to transfer this rule where it is not applicable, thus indicating that they did not understand the implications of the LLN.

Materials

Training Materials

Four training conditions were included in the study. (All of the training materials are listed in Appendix A.) The training conditions were as follows:

(1) Minimal training condition;

(2) Rule repeated condition;

(3) Expanded explanation condition;

(4) Full training condition.

Because it has been shown that some training is better than no training (Fong et al. 1986; Well et al. 1990), the training conditions were not tested against a no training condition. Rather, a "minimum training" condition was included in order to ascertain whether subjects merely learn the statement, "bigger samples are more accurate" from training. If the subjects in this group did not do significantly worse than the subjects in any of the other training conditions, this would suggest that the subjects in all conditions might be only remembering this most basic statement and have only an extremely superficial understanding of the LLN. The demand condition in the Fong et al. (1986) study in which subjects read a one-sentence introduction to the LLN followed by an example might be thought to be similar to the minimum training condition in the present study. However, it is difficult to ascertain exactly what the subjects read in the Fong et al. (1986) study. Subjects in the Fong et al. demand condition reportedly

produced higher quality answers that were classified as being statistical more frequently than subjects in the control group. Although the improvements were not significant, it was deemed worthwhile to test this type of condition again.

In the training materials used by Fong et al. (1986) and Fong and Nisbett (1991), simplistic definitions of the LLN were repeated several times. For example, in the rule training condition in the former study, a definition of LLN was given seven times. It is possible that subjects simply memorized a definition of the LLN without really abstracting the concept. Therefore, in the second condition used (rule repeated), subjects received a written explanation of the rule. In this short statement the LLN was repeatedly defined in terms taken from the Fong et al. (1986) rule training condition. A statement indicating that the LLN is important when drawing samples was also included. If subjects reading this statement were able to do as well as the subjects in the other two training conditions, their performance may indicate that subjects do not understand and abstract the rule in a manner that enables them to apply it across domains. Rather, they are able to repeat what they have been told is important without understanding the concept.

The third training condition is similar to a condition used by Well et al. (1992) in recent research. In one study a written explanation condition was included in order to determine whether it was as helpful as a full training condition. Although students did better on problems after receiving this explanation than they did with no training, (38% correct vs. 23% correct in immediate

testing) they did not do nearly as well as they did in the full training condition (75% correct in immediate testing). One confounding variable Well et al. later discovered was that in the full training condition subjects received a LLN problem (the post office problem, Appendix A) before the training. In the written explanation condition subjects did not receive this problem. Because subjects may have been more interested in the training after receiving a problem, Well et al. decided to test a condition in which the same LLN problem was given to subjects before they received the written explanation. The results indicate that presenting this problem did in fact improve performance on subsequent problems (58% correct vs 35% correct in the written explanation condition).

Therefore, the expanded explanation condition was included in this study in order to obtain some knowledge about what subjects learned in this condition. It is possible that they did abstract an understanding of the LLN after being primed by a problem where the rule was applicable. It is important to know whether subjects learn as much about the LLN in this condition as subjects in the Fong et al. (1986) full training condition, or as much as subjects in the rule repeating condition. Subjects in the Fong et al. (1986) full training condition received example problems after receiving the written and verbal explanations of the LLN. It is pedagogically important to ascertain whether making subjects aware of the applicability of a rule by introducing a problem before training encourages students' learning more than explaining examples after training.

The last training condition was nearly identical to the Fong et al. (1986) full training condition. The rule training and the guided induction training that comprise this condition were taken from the Fong et al. (1986) materials. There were only two differences between Fong et al.'s (1986) full training and the full training condition in this study. The first is that black and white marbles were used in this study instead of red and blue gumballs. The second difference is that only two examples were explained instead of three. This change was made due to time constraints.

If after training in this condition subjects are not only able to give correct answers more frequently, but also to provide more sophisticated rationales than subjects in the other conditions, the full training can be considered unique and valuable.

In summary, the purpose of all of the above conditions was to ascertain whether subjects were able to abstract the LLN and apply it across several problem types without being directly told to do so in any of the conditions. Additionally, if they were able to, which training conditions lead to this, and what did subjects abstract?

Testing Materials

The different aspects of the LLN that were to be studied in each of the problems were variability and accuracy. In order to learn the most about what subjects understand, several kinds of problems were used including some open-ended problems from Fong et al.'s study (1986) as well as some multiple choice problems used by Kahneman and Tversky (1972) and Well et al. (1990). All

of the problems are listed in Appendix B. In the immediate testing condition subjects received eight problems that dealt with some aspect of the LLN and one filler problem.

The eight problems included two open ended problems (taken from Fong et al. 1986), two lure problems, one false alarm (taken from Fong et al. 1986), one center version problem and two tail version problems. The center version and tail version problems were described and used by Well et al. (1990). Center version problems were stated in terms of the center of a sampling distribution whereas tail version problems were stated in terms of the tails of a sampling distribution.

The purpose behind selecting these eight problems was to present subjects with a variety of problems which allowed them to give answers that rely on the concepts of accuracy and/or variability, while also allowing them to show that they would not overapply the LLN where it was not applicable. The variety of problems also allowed for a comparison between subjects from each training condition on different types of problems. The false alarms and lures were intended to test the students' knowledge about an aspect of the LLN that does not simply follow the "bigger is better" rule. Two lures were included and only one false alarm because the lures provide a more stringent test for understanding the LLN than the false alarms do. These three problems combined with the filler made up half of the problems subjects were given in the first testing session. This ensured that subjects were given ample opportunity to realize that the LLN may not apply to all problems.

Also, Well et al. (1990) found that subjects were significantly better able to correctly answer center version problems than tail version problems. They suggested that this was due to the fact that while subjects may understand the increased accuracy of a larger sample's mean, they frequently still did not understand the the effects of sample size on variability. Therefore, because they more directly lend themselves to explanations based on the idea of variability than do the center version or Fong et al. (1986) problems, two tail version problems were included.

Because no training time was required in the second session, more time was available for problems, therefore two additional problems (another center version problem and another open ended problem) were included. The additional open ended problem was intended to act as another false alarm, however the problems selected were later determined to be extremely different from the false alarms and were therefore thrown out. The problems used were counterbalanced within groups across testing sessions.

Method

Subjects

The subjects were recruited from the University of Massachusetts Department of Psychology subject pool. They consisted of 48 undergraduates who received class credit for their participation. None of the subjects had previously taken a college

statistics course. The subjects were divided into training groups of 1-4 students.

Procedure

Within each small training group all subjects received the same training materials. The subjects were given the training materials to read at the beginning of the session. When the subjects were finished with the training materials, the testing problems were administered. Originally these were going to be given one-by-one to the entire group (that is, a problem would be passed out to all subjects only when the previous one had been completed by all of the subjects) in order to encourage the subjects to take the time to write an explanation down, rather than racing through the problems. However, there was an extremely large amount of variability in the speed at which the subjects worked. Therefore, each subject was given all of his/her problems at one time and was given the following instructions,

> "I am going to be giving you several problems to solve. I would like for you to give each problem your best answer. However, I do not want you to be worried if you are not sure about which is the correct answer. Give me your best answer. I am more interested in why you gave the answer that you gave for each question than whether or not you get the correct answer. Therefore, after each question you will be asked to give your rationale for the answer, or to explain what you think about the rationale given in the problem. Please take your time to do your best at explaining your reasoning, even if you are not sure the answer is correct. Please take your time to do this. You will have plenty of time to finish. I would like you to do one problem at a time. Do them in the

order in which they have been given to you. After you finish one please turn it over and continue with the next problem. Do not worry about going back to any questions. If you have any questions at any time, please feel free to me. Remember to take your time, you have plenty of time."

The subjects in the control group were given the one sentence definition (see Appendix A), the above instructions and then the problems. The subjects in the expanded explanation group were given the post office problem to solve. When the subjects were finished with the problem they were asked to carefully read the explanation. After all of the subjects were finished with the explanation, they were given the above instructions and the problems were administered. Likewise, subjects in the rule repeated group were asked to carefully read their explanation. When all of the subjects were finished reading, they were given the instructions followed by the problems. The full training group was initially asked to carefully read the written abstract explanation. When all subjects were finished reading, the brief oral presentation of the LLN was given. The full training group was then given the explained examples. Subjects were asked to carefully read and answer each problem and then to carefully read the explanation proceeding the problem. They were then given the previously mentioned instructions followed by the testing problems. The entire session for the full training group lasted about an hour. Sessions for each of the other groups lasted about 35 to 45 minutes.

The second testing session occurred approximately two weeks after the first (mean=14.81 days, standard deviation=1.97). All subjects received the same instructions as in the first session. The second session lasted 30 to 40 minutes.

Half of the subjects in each condition received one set of problems in the first session and the other set of problems in the second session. The order of problem sets was counterbalanced across conditions. The order of the problems remained the same for all subjects with the starting point in the order randomized.

Analysis and Results

Three separate coding systems were used to score the answers and the written rationales for them for each question. Scoring was based on the following three criteria:

(1) whether or not an answer was correct;

(2) the three-point coding system developed by Fong and Nisbett (Fong et al. 1986; Fong & Nisbett 1991); and

(3) a qualitative coding of the rationales.

Each of these analyses will first be reported separately, followed by any appropriate comparisons between analyses.

Analysis of Correctness

Each correct answer was given a score of 1, and each incorrect answer was given a score of 0, regardless of the rationale given after the answer. In the multiple choice problems it was

simple to determine whether subjects chose the correct alternative. However, the scoring for the open-ended problems was not so straightforward. The answers for the false alarm problems were given a score of 1 if there was no indication that a larger sample was needed. The open-ended test problems were given a 1 if the subject correctly agreed or disagreed with the rationale given in the problem. It is important to emphasize that scoring a problem as correct did not mean that the rationales the subjects gave for agreeing or disagreeing were necessarily correct.

Six totals were obtained for each subject. These included the total correct for each of the following groups of questions for each testing session: (1) Test problems, open-ended problems (excluding the false alarms), as well as the center and tail version problems; (2) lures; and (3) false alarms. These totals were not combined because the problem types were considered to be too different, and the individual totals were initially of more interest. If subjects in one training condition were able to correctly answer all test questions, while subjects in another condition were only able to answer the lures and false alarms, the results would indicate a great deal about what subjects learned.

The mean proportions correct and standard errors of the means for each of the training condition's totals are listed in Table 1. Overall, subjects' performance was quite poor. The across group averages indicate that subjects correctly answered less than 50% of the test and lure problems and about 70% of the false alarm problems. Training group did not have an effect on performance, in fact the average percent of correct test questions for groups 1, 2,

Table 1

Average Scores for Problem Types by Groups for Each of the Two Test Sessions

Training Group	Problem Type					
	Test		Lure		False Alarm	
	first	second	first	second	first	second
Minimum	.556	.396	.375	.375	.917	.583
	(.075)	(.065)	(.125)	(.125)	(.083)	(.149)
Rule Repeating	.500	.438	.500	.250	.750	.667
	(.112)	(.070)	(.087)	(.115)	(.131)	(.142)
Expanded Explan.	.472	.521	.500	.625	.667	.667
	(.104)	(.104)	(.138)	(.125)	(.142)	(.142)
Full	.556	.479	.292	.458	.667	.583
	(.063)	(.084)	(.114)	(.130)	(.142)	(.149)
Average	.521	.458	.417	.427	.750	.625

Note: Mean scores based on 0,1 coding. The standard errors of the means are listed in parentheses.

3, and 4 were 48%, 47%, 50% and 52% respectively. Analyses of variance conducted for group differences with independent variables question type, and testing time did not reveal any significant differences between training groups. However, the mean proportions correct indicate that all subjects performed a little better on false alarm questions than on test or lure problems. A more detailed item analysis will be discussed later. Additionally, no subjects appeared to perform well on any types of problems. The one exception may be the minimal training group's and the rule repeated group's performance on the first session false alarm questions. However, the differences in performance levels were not significant.

Fong and Nisbett Three-point Scale Analysis

An analysis based on the Fong et al. (1986; Fong & Nisbett 1991) three-point coding scheme was conducted. This was done in order to ascertain whether or not the present results were actually in conflict with the Fong et al. results, or merely different because the scoring system was different. Fong et al. describe this coding in the following manner:

(1) A one was given for an entirely deterministic answer.

(2) A two was given for a poor statistical answer.

(3) A three was given for a good statistical answer.

A more detailed description of the three-point coding scheme scorers used is provided in Appendix C. In order to assess reliability, one-fourth of the data were randomly selected and coded by the author and one other person. There was exact

Table 2

Average Scores Based on Fong and Nisbett Three-point Coding System for Two Testing Sessions

Training Group	Problem Type					
	Test		Lure		FA	
	first	second	first	second	first	second
Minimum	1.861	1.750	1.583	1.583	1.583	1.583
	(.112)	(.138)	(.120)	(.104)	(.229)	(.260)
Rule Repeating	1.944	1.917	1.750	1.792	1.500	1.417
	(.178)	(.167)	(.131)	(.096)	(.195)	(.193)
Expanded Explan.	1.833	1.938	1.917	1.792	1.167	1.583
	(.195)	(.213)	(.120)	(.130)	(.112)	(.229)
Full	2.056	1.938	1.917	1.792	1.833	1.833
	(.122)	(.138)	(.120)	(.130)	(.208)	(.167)
Average	1.924	1.885	1.792	1.740	1.521	1.604

Note: Maximum possible score=3. Mean scores are followed by the standard errors of the means in parenthesis.

agreement for 75% of the problems. Table 2 shows the mean scores and standard errors according to the three-point coding by group, for each of the previously discussed totals.

Once again, the overall level of performance did not appear to be good. All of the average scores were less than 2, with 1.924 as the highest. Averaged across testing times the averages for test, lure, and false alarm problems were approximately 1.9, 1.8 and 1.5 respectively. When analysis of variance tests were performed, no significant differences were found between groups for test problem performance in either the first or second session. Although there was some tendency for the scores to be higher for the full training condition (group 4) than the others, e.g. test question average 2.04 as opposed to 1.81, 1.93, and 1.89 for groups 1, 2 and 3 respectively, the estimated effect sizes were small. Given the estimated effect sizes and variances, power calculations indicated that in order to have a power of .9 for the group main effect, sample sizes of 83 and 50 would be required for the lure and false alarm problems respectively. The estimated effect size for the test problems was zero. It is important to note that the average scores were all approximately between 1.5 and 2.0, with one exception being the expanded explanation's average score for the false alarms in the first testing session. Thus on average, subjects performed just below the "poor statistical reasons" level. The lack of differences between groups is particularly obvious when looking at average scores in the second

testing session for groups 2, 3 and 4 on test problems and lure problems.

No significant differences were found on levels of performance between the first and second testing sessions for any of the groups. Additionally, no significant effects were found for question order.

Analysis of Rationales

In addition to the quantitative analyses, qualitative analyses of the rationales subjects gave for their answers were also performed. Subjects' answers were coded as being consistent with one of twenty-two rationales. The twenty-two rationales were divided into seven broad categories which are listed and briefly described below.

1) No regard for sample size. That is, sample size is mentioned and discounted, or any sample is said to be good.

2) Bigger is somehow better. This includes the extremely simplistic statements that say nothing more than bigger is better.

3) No regard for data. Generally, this included statements indicating information other than the given data was more important in analyzing the questions than the data.

4) Small samples are inadequate. Rationales in this category included pointing out some problem with using a small sample to make conclusions about a population.

5) Big samples are adequate. Rationales including some statement about why big samples are adequate.

6) Some comparison between samples. These rationales generally indicated the deepest understanding of the importance of sample size.

7) Other. Generally these were somewhat nonsensical, restatements of the problem, or otherwise virtually impossible to classify.

The seven categories and twenty-two rationales which were used for the coding were generated by looking at subjects rationales and attempting to create a coding system which would accurately describe the differences in the types of answers received. Table 3 and the subsequent analyses are based on the seven broad categories. However, it is informative to look at the more specific types of rationales which were included in each category (See Appendix D). The purpose of coding the results in this manner was to prevent false indications of understanding the LLN. It was also to obtain the maximum amount of information about what subjects do understand and are able to express compared to what concepts they do not seem to understand. Two scorers each scored one-fourth of the data and produced 77% agreement in the categorical coding, one of those scorers also scored the remaining data.

Table 3 provides frequency distributions for the categories of rationales used for each type of problem, collapsed across testing time. Approximately 5% of the rationales given could not be classified as being part of only one category, and were therefore included in both of the appropriate categories. Because each testing group had different total numbers of rationales, the distributions are reported in percentages.

Table 3

Frequency Distributions of the Rationale Categories Used by Group, Collapsed Across Testing Times

Type of Question	Group	Rationale Category 1	2	3	4	5	6	7
Test	1	6.8	23.9	28.4	17.0	1.1	12.5	10.2
	2	8.0	18.2	14.8	25.0	0.0	19.3	14.8
	3	13.0	7.6	13.0	23.9	2.2	22.8	17.4
	4	9.2	5.8	17.2	20.7	0.0	37.9	9.2
Lures	1	7.6	41.5	15.1	7.6	3.8	11.3	13.2
	2	8.3	37.5	8.3	4.2	2.1	25.0	14.6
	3	24.1	13.0	11.1	13.0	9.3	22.2	7.4
	4	20.0	10.0	12.0	6.0	0.0	40.0	12.0
False Alarms	1	4.2	0.0	58.3	0.0	25.0	8.3	4.2
	2	0.0	0.0	70.8	4.2	16.7	4.2	4.2
	3	0.0	0.0	50.0	8.3	25.0	4.2	12.5
	4	0.0	0.0	29.2	12.5	37.5	20.8	0.0

Note: The numbers listed are percentages of the total number of rationales for that row.

Differences were found in the kinds of rationales that subjects who received different kinds of training provided. For example, subjects in group four (full training) consistently had the highest percentage of rationales from category 6, which includes most sophisticated rationales, those which comparing the sample sizes in the problems. The percentage of category 6 rationales for the full training group was 39.0 for test and lure problems combined, and 20.8 for false alarm problems . The corresponding figures for the minimum, rule repeating and expanded explanation training groups were 11.9 and 8.3, 22.2 and 4.2, and 22.5 and 4.2 respectively. Additionally, the minimum training group had the highest percentage of rationales from category 2, a simple version of "bigger is better". The minimum training group had 23.9% of test and 41.5% of lure problem rationales from category 2, whereas the percent of rationales from category 2 for the full training group were 5.8% for test problems and 10% for lure problems. Another important trend to notice in Table 3 is that groups 1 and 2 performed similarly to each other, but different than groups 3 and 4, which also performed quite similarly. Collapsing the percentages from groups 1 and 2 and groups 3 and 4, the resulting percentages for test questions rationales in category 2 become 21% compared to 6.7%, and for category 6, 15.9% versus 30.3% respectively. Although subjects from different training conditions did not perform significantly different on problems, it appears that they relied on different rationales to answer the problems.

In order to test whether the mentioned differences in the types of rationales subjects used were significant, the previous

categorical scores were divided into "good" statistical reasons and "poor" statistical reasons. For test questions, rationales from categories 4, 5, and 6 were considered to be "good" and were given a score of 1, while all other rationales were given a score of 0. Although rationales from category 1 might have been more appropriate for lure problems than those from categories 4, 5, and 6, the lures and false alarms were also given the 1,0 coding in order to see if subjects were using the more sophisticated rationales on all problems, including those that did not necessarily require their use. Additionally if subjects used the more simple or the more sophisticated rationales on all kinds of problems, it is interesting to know whether the use of the rationales lead to correct or incorrect answers. Average scores from the 1,0 rationales coding were obtained for test questions and false alarms/lures (FA/L) for each subject. Using this system, the maximum possible score would be 7 for the test questions and 13 overall. The means and standard errors are listed in Table 4. These means indicate that none of the groups performed exceptionally well. The full training group (group four) received the highest scores overall, however, their means are only 3.667 and 3.350 for the test questions and the false alarm/lures respectively. Across groups the mean scores are all less than half of the maximum possible. The scores for group four were found to be significantly higher than those for group one (minimal training) on both the test questions and the FA/L problems ($F(1,44)=5.253$, $p<.05$; $F(1,44)=5.031$ $p<.05$). Significant differences between groups

Table 4

Mean Number of Sophisticated Rationales Used by Group for Test and False Alarm(FA)/Lure Problems

Training Group	Questions		
	Test	FA/Lure	Total
Minimum	1.833 (.366)	1.583 (.484)	3.417
Rule Repeated	2.833 (.626)	1.833 (.534)	4.667
Expanded Explan.	2.917 (.763)	2.417 (.570)	5.333
Full	3.667 (.414)	3.250 (.509)	6.917
Average	2.813	2.271	5.083

Note: Each subject received 7 test questions and 6 FA/Lure questions. These scores are based on a 1,0 scale, in which subjects received a 1 for an answer which belonged to category 4, 5, or 6.

three and four compared to groups one and two were also found for the false alarm and lure problems ($F(1,44)=4.585$, $p<.05$).

Other Results

There are a few other results that should be mentioned. First of all, none of the analyses indicate significant differences between the first and second testing sessions. Secondly, subjects in all groups performed quite differently on various questions. Table 5 shows the mean scores across groups for each question according to the 1=correct, 0=incorrect coding. It is interesting to note how well subjects performed on the open-ended questions as compared to the other test questions. For example, the average score across groups for the open-ended test questions was approximately .69, whereas the average for the multiple choice test questions was approximately .38. Likewise, the average across groups on false alarm questions was approximately .69, but the average for lure questions is approximately .42. These differences are particularly interesting when one considers the previous research done on understanding the LLN. Kahneman and Tversky (1972; Tversky & Kahneman 1971, 1974) have consistently used multiple choice problems and concluded that people do not understand the LLN well, yet Fong and Nisbett (1991; Fong et al. 1986) have used open-ended problems and concluded that people can learn and use the LLN appropriately. One inherent difference in the problems types is the chance level of performance. For the open-ended problems random guessing

Table 5

Mean Scores for Each Problem by Group

	Group				
Question	Min.	Rule	Expan.	Full	Avg.
Open Ended					
Test					
IRS	.750	.833	.833	.750	.792
Slot	.583	.583	.583	.583	.583
False Alarms					
Brew	.917	.917	.750	.667	.813
Audit	.583	.500	.583	.583	.563
Multiple Choice					
Lures					
Poll	.333	.500	.667	.583	.521
Fish	.333	.333	.333	.333	.333
Pers	.333	.250	.667	.417	.417
T.Dept.	.500	.417	.583	.167	.417
Center					
Book	.750	.333	.417	.583	.521
Tennis	.583	.500	.583	.500	.542
Tail					
Blood	.250	.333	.333	.333	.313
Hosp	.333	.417	.417	.417	.396
Women	.000	.167	.333	.333	.208
Geol	.000	.333	.333	.500	.292

Note: Mean scores for each question based on 0,1 coding in which subjects received a 1 for a correct score. The problems are all listed in Appendix B.

would yield a performance level of.5, but for the multiple choice questions random guessing would yield .33.

It is also important to notice the differences in performance on the center and tail versions of the multiple choice questions with averages across groups being .53 and .30. This difference is consistent with the differences obtained by Well et al. (1990). However, even within problem type there were differences in the way subjects performed on different questions. Subjects' overall average on the IRS problem (.79) was much better than the average on the slot machine problem (.58) and likewise, better on the brewery problem (.81) than on the auditor problem (.56). On multiple choice questions subjects performed better on the pollution problem (.52) than on the fish hatchery problem (.33), and better on the hospital problem (.39) than on the women's heights problem (.21). The manner in which a question was worded made a big difference in subjects' performance levels.

One of the most interesting questions that arises from the analyses that were conducted concerns the extent to which the performance levels on the various coding schemes correlate. The correlations between the number of correct answers and the number of sophisticated rationales for each subject by group were calculated in order to gain understanding as to whether or not subjects do give correct answers with poor rationales and wrong answers with sophisticated rationales (See Table 6). Due to large standard errors there are no interesting significant differences between any of the correlation coefficients even though some of the coefficients appear to be quite different from others. In

Table 6

Correlation Coefficients for Number of Correct Answers and Number of Sophisticated Answers

Group	Correlations for Test Problems		
	r_{cf}	r_{cr}	r_{fr}
1	.747	.723	.793
2	.856	.833	.888
3	.402	.902	.442
4	.847	.630	.803

Note: r_{cf}=correlation coefficient for correctness score and Fong et al. coding score

r_{cr}=correlation coefficient for correctness score and rationales score (avg. number of scores 4, 5 and 6s)

r_{fr}=correlation coefficient for Fong et al. coding scores and rationales score

particular, the expanded explanation group's coefficients for the correlations between the correct score and the Fong et al. coding and between the Fong et al. coding and the quality of rationales, stand out because they are so much smaller than the other test question correlations. Scatterplots of these relationships indicate the expanded explanation group's low r_{cf} for test questions appears to be mainly due to one outlier and the low r_{fr} for test questions appears to be due to the fact that the data are in two clusters.

The correlations were also calculated for the false alarm/lure problems, however, due to the fact that the rationales coding for these types of problems is not necessarily indicative of understanding the LLN correctly, these coefficients have been omitted. However, as one would expect, many of the correlations for r_{cf} and r_{cr} are negative.

CHAPTER III

DISCUSSION

The present results do not seem to support the results of Fong et al. (1986) and Fong and Nisbett (1991) about the effectiveness of training. The analysis conducted for the coding scheme used by Fong et al. (1986) and Fong and Nisbett (1991) indicates that subjects may not in fact be learning all that Fong et al. (1986) and Fong and Nisbett (1991) claimed they were learning. No significant differences in between-group performance were found in the present study. However, the mean scores found using the three-point coding scheme were similar to those reported by Fong and Nisbett (1991). With the exception of the full training group in the present study, the means for all of the conditions in both studies were between 1.5 and 2.0. Although the group means were found to be significantly different in the Fong and Nisbett (1991) study, power calculations indicate that much larger groups would be needed to find significant differences between the groups in a replication of the present study for the false alarm and lure problems (50 and 83 subjects per condition respectively, as opposed to the 12 subjects per condition that we used). Even though the group means for the test questions suggested an advantage for the more extensive training conditions, the analysis of variance F for the condition main effect was less than 1, suggesting an effect size equal to 0. This information suggests two things. First of all, despite the fact that the effects Fong et al.

report were statistically significant, these effects were in fact quite small, and statistical significance was found because of the large sample sizes and numerous problems given. Additionally, it is apparent that the performance of subjects in the present study and in Fong et al.'s studies were poor, with scores on the average below 2.00, which was the score given for a "poor" statistical reason.

The analysis based on the number of correct answers also indicated poor performance by subjects. Not only were there no significant differences between groups for this measure, but also the estimated effect sizes for test questions, false alarms and lures were all equal to 0. None of the groups of subjects performed very well. In Table 1 we see that the mean scores for each problem type were less than or equal to .5 for one half of the overall averages. Average proportion correct for first and second testings of the lure problems across groups was .42 and .43 respectively, barely better than the .33 that would be obtained by randomly choosing an answer. For the open-ended problems (test and false alarms) chance performance was .5; that is, in order to receive a score of 1 the subject merely needed to correctly agree or disagree with a stated conclusion. Therefore, chance performance for testing questions (two open ended and five multiple choice) would be equal to .38. Across groups and testing times the subjects' average score was equal to .46. Subjects did perform much better on the false alarm problems, with the overall average being approximately equal to .69.

Few subjects in any of the groups performed well on all of the problems. No subjects correctly answered every problem. Only two subjects (one from group 2, one from group 3) correctly answered all of the test questions, and only two subjects (both from group 4) correctly answered all of the false alarms and lures. Four subjects (two from group 3, two from group four) missed one test question, and four subjects (one from group 1, one from group 2, two from group 3) missed one false alarm/lure. Only three of the 48 subjects in the entire study missed one or less on both test questions and false alarms/lures.

Although no significant differences were found in the quantitative analyses, the story changes a little bit when we look at the qualitative analysis of rationales. Subjects in the advanced training groups may not have provided any more correct answers than the subjects with minimal training, but, they did attempt to use more sophisticated rationales. Looking at the frequency distributions from Table 3, we see that subjects in the full training condition were more likely than any of the other training groups to use rationales from category 6, which is considered to contain the most advanced rationales (37.93% vs. 18.22% average on test questions for groups 1, 2, and 3). However, the full training group had the lowest correlation between number of correct answers and number of sophisticated rationales used for test questions (see Table 6). This suggests that subjects attempted to use the knowledge obtained in the training sessions, but did not know how to do so. This evidence clearly conflicts with the formalist theory. That is, subjects did not, as the formalist theory would suggest,

learn an abstract rule and properly apply it across domains. Rather, they appear to have learned the key words without understanding the concepts any better than subjects who received less training.

Given that subjects do not adequately learn how to correctly apply the LLN to the problems they are given in the brief training sessions, the question becomes, what exactly were the subjects learning? Looking at Table 3, we see that subjects from different groups gave different kinds of rationales. The subjects in the minimum training and rule repeated conditions tended to use rationale 2 (bigger samples are...), much more frequently than subjects in the expanded explanation and full training conditions (21.02% vs. 6.68% on test questions; 39.55% vs. 11.48% on lures). Rationales in this category include extremely simplistic, not applicable, or generally wrong statements about big samples such as, "Bigger samples give greater chances for getting anything", "bigger samples give bigger percentages", etc. These are the types of statements that subjects might have learned from the minimum training and rule repeated conditions, or already knew before they received any training. However, subjects in the expanded explanation and full training conditions seemed to learn more complex statements about the LLN, and regardless of how well they understood the statements, they attempted to use them. Although some people may consider this memorization learning, these subjects certainly did not learn in the sense of understanding, rather they simply remembered key phrases.

As was mentioned in the introduction, it is extremely important to define exactly what is meant by understanding the LLN. For a statistician this probably means understanding why all of the statistical analyses one does can be done, and the implications of sample size on all of the statistics the analyses produce. However, for the purposes of this training study, we can consider understanding the LLN to be understanding the ideas of accuracy and variability and the effects of sample size on these concepts. For example, simply saying "A bigger sample is better", would not be indicative of true understanding of the LLN. Although it is possible that a subject who provided such an answer may have a thorough understanding of the LLN, this does not seem likely for a number of reasons. One reason is that if the subject did truly understand the LLN, he or she should be able to answer most or all of the problems correctly. No subjects in this study correctly answered every problem. Only three of the 48 subjects in the entire study missed one or less on both test questions and false alarms/lures. This indicates that perhaps three subjects had an adequate understanding of the LLN.

It is also unlikely that those providing only simplistic rationales had a good understanding of the LLN because the subjects were strongly encouraged to explain as well as possible why they choose the answer that they did. Most subjects took the time to write quite a bit for their rationales. Subjects who gave an answer such as "bigger samples are better" generally provided more words that either did not say anything more, or were irrelevant. When subjects did know more, they usually wrote

more, providing answers like those listed in category 6 (see Appendix D). However, as the correlation coefficients show (see Table 6) subjects who used rationales from category 6 did not always provide correct answers. It is encouraging to note that the correlations between the number of times sophisticated rationales were used and the number of correct answers on test questions for groups the minimal, rule repeated, and expanded explanation training groups were all quite high, .723, .833, and .902 respectively. However, due to the fact that the correlation coefficient for the full training group (group 4) was only .630, the coefficients are difficult to interpret.

It appears that the subjects in the first three groups that showed sophisticated reasoning knew how to apply this reasoning to test questions. The high correlation for the expanded explanation group might be due to the fact that those subjects were required to attempt to answer a question similar to the more difficult testing questions at the beginning of their training sessions. Therefore, those subjects that were interested in learning how to correctly answer the problem were primed to learn the reasoning behind the correct answer. However, it is difficult to understand why the correlation coefficients were higher for the minimum training and rule repeated conditions than they were for the full training condition. Perhaps the few subjects in the former two groups who did perform well had some knowledge about the LLN prior to the study, although none of the subjects had previously had a statistics course. The subjects in the full training condition, like those in the expanded explanation condition, also

received example problems in their training. However, their example problems were more like the "easier" testing problems than the more "difficult" problems (based on how subjects performed overall). Although they may have then been better able to answer problems like the examples, this knowledge may not have transferred to the more difficult multiple choice questions. Their example problems were also given at the end of the training, rather than at the beginning. Thus they could not have been primed to learn how to answer probabilistic questions at the beginning of their training.

The analysis of the rationales not only provided information about what subjects learned in the training conditions, it also inadvertently provided information about some basic concepts with which students have difficulty. One somewhat surprising finding was the lack of understanding of basic mathematical concepts, such as percentages or proportions. For example, many subjects continually provided rationales indicating that bigger samples give bigger percentages. This rationale did in fact lead many of the subjects to correctly answer some of the questions. However, it is an indication of a lack of understanding of percentages. Given their lack of basic mathematical understanding, it is not surprising that many undergraduates have difficulties understanding probabilistic and statistical rules.

It is important to understand the implications of the present study for the empiricist/formalist debate. Although the current results do not directly support either position, they seem to contradict the formalist theory in particular. Subjects did not

correctly apply what they were taught in training, nor did they appear to rely on an abstract rule. Although the current study was lacking in statistical power, the estimated effect sizes for analyses of the test questions based on both of the quantitative coding schemes were zero. This indicates that minimal training was as effective as any of the training employed. However, looking at the analysis of rationales, we see that there are discrepancies in the types of rationales on which subjects rely. This is the point at which the empirical and formalist debate becomes confusing. After training, subjects appear to attempt to use the rules that they were taught, by providing simplistic rationales with words similar to those used in training. However, they do not do this very well. In fact, this effort tends to lead many of them to the incorrect answers. Therefore, one can not say that subjects are correctly using an abstract inferential rule. One can only say that subjects are attempting to use rules that they do not understand.

Does this then support the empirical point of view? The results do not strongly support the view. However, it seems that subjects learned a specific rule and attempted to use it. The rules they used were not always appropriate, so in some sense they did attempt to use a domain specific rule. However, with a better understanding of the rules, subjects might have been able to use the rule correctly. However, trained subjects do not perform any worse on the false alarms and lures than the subjects who were not trained, indicating they do not necessarily over-apply the rule. Yet they also do not perform any better on the test questions, indicating that maybe they did not really learn how to use the

LLN. These results make it difficult to discern exactly what subjects learned from training. We can look at the qualitative results and find the words that subjects have learned, but that brings us back to the question, what is learning? Being able to recall the words that describe a concept may be considered by some to be learning. In effect, it is learning; learning words. However, it is not learning the concept. In order to learn a concept one must go beyond memorizing words to understanding ideas.

At what point do people make the leap from memorization to understanding? Although this question is extremely difficult to answer, several people have attempted to answer it. Holland et al. (1986) consider knowledge to be represented in terms of hierarchical rule structures. According to their theory, we reason based on a rule hierarchy built through induction. In this hierarchy we store a basic default value that we generally use to solve problems. These default rules are usually predictive and are therefore followed in most circumstances. However, in certain instances these default rules are overridden by specific level exception categories and rules. According to Holland et al. (1986), the default hierarchy is a way of representing generalizations that can be used to model the world.

People may attempt to reason statistically by depending on a default hierarchy similar to that referred to in the inductive reasoning literature (Holland et al., 1986). Rather than transferring abstractly across domains, or only reasoning within the given domain, perhaps people simply use a default value that is always available to them. This is not to say that default values

do not change with training. It also does not mean that the hierarchy has to change with training. People will rely on the most basic level of knowledge that they have that seems to work.

The simplistic form of the LLN, "bigger is better", (when taking samples from a population) has been referred to as a statistical concept people learn early in life (Nisbett et al. 1983). This may be learned through experience in several areas of life, and used without understanding the more detailed and technical aspects of the LLN. Thus, early in life a default value may fall into place which is something like, "a small sample is not necessarily indicative of a large population, rather a large sample is better". When people experience events that require an explanation, they may try to understand it by falling back on this basic reasoning. If the LLN were placed in a hierarchy people may have "bigger is better when samples are taken" as their default value. Unless this hierarchy for the LLN is further built upon, people will use this basic reasoning in any situation in which it seems applicable. This is true even if the event should not be explained by such a rule but a person can find no better rule to apply.

An important part of the LLN lies in the fact that it is not one concept. Rather, it consists of many subconcepts each of which can be learned and explained without reference to the LLN. These concepts include accuracy, variability, regression to the mean, sample distributions, sampling distributions, population distributions, estimation, confidence intervals, and others. However, all of these subconcepts are strongly tied to the overall concept. For example, it is quite possible to answer a LLN question

with "bigger is better" without knowing about what regression to the mean is, let alone knowing that when taking samples, extreme averages will eventually be balanced by averages that approach the population mean. While it may not be necessary for all subjects to understand all statistical concepts in order to get the gist of the LLN, if "bigger is better" is all that they know, it should not be claimed that they have correctly obtained the abstract rule.

However, when people acquire more training in statistics, they build more branches off of their main default level. Not only do they then have more branches which can be used to access specific concepts within the LLN, such as regression to the mean, but these branches also become more complex. That is, the definitions for each subconcept become more complex. Thus, the expert is able to identify specific exceptions to the basic rules on which the novices rely. That is not to say that the experts will not use the default rules; on the contrary, they will remain the most frequent and standard explanations. Yet, the expert has additional knowledge from which can be derived more accurate answers to problems that need an explanation beyond the basic default rule. While to the expert it may seem that each aspect that is a part of the overall concept of the LLN is intimately tied to the others, I do not think that this is how the novice thinks.

If each aspect of the LLN has its own branch off of the default value it would be possible to correctly answer many probabilistic questions after some training, with only a partial understanding of the overall concept. I think that this may be what has happened to subjects in statistical training studies. The

training builds upon the basic default value rule the subjects already have. One could call this default rule an extremely basic abstract rule. While this is better than no knowledge, it may not be accurate knowledge, rather, it may be lacking in important concepts, and overapplied where it is not necessary. However, I do feel that with training people can be taught not to overapply their default rule, perhaps by being shown examples in which their default rule does not apply. They can also be taught that the rule is a simple way of expressing several statistical concepts, and that they should be aware of the fact that exceptions to that rule exist.

In the present study, subjects with minimal training probably either used the default rule that they already had, or attempted to use the simplistic rule they were given in training as their default rule. Therefore, they were able to answer some of the questions correctly, but did not possess a knowledge structure adequate to correctly answer all of the questions nor to provide sophisticated rationales for doing so. However, the more advanced training groups probably attempted to either change their default rule to something more complex than they previously had, or to extend their already existing default rule without truly understanding what the new words they had been taught meant in terms of what they already knew. Thus, they probably did not know how to correctly use their new knowledge. Perhaps this was because they thought that it now applied to more situations than they had previously been aware, or they thought that it no longer applied to the situations for which their old rule had been appropriate.

It appears that it is not possible to teach subjects a great deal about the LLN in brief training sessions. Educators must be aware of the fact that the concepts are not so simplistic that they can be quickly learned and readily applied by students. Perhaps the wisest thing that educators could do would be to attempt to teach students one part of the LLN at a time, and then after students have learned something about that subconcept only to claim that their students understand that specific concept, not that they have an understanding of the LLN. We must be cautious in claiming that students understand concepts for which they only have a simplistic default rule. Once we have claimed that they understand a concept, we might close the door to needed further education that we assume is auxiliary. It is imperative that we not only teach students important probabilistic concepts, but also that we show them how to apply these concepts to various situations. Students' hierarchical knowledge structures must not only contain rules, but also levels which include information about practical applications of the rules contained within the hierarchy.

APPENDIX A

TRAINING MATERIALS

1. CONTROL GROUP

The Law of Large Numbers is a rule that states that as the size of a random sample increases, its distribution becomes a more accurate estimate of the distribution of the population. That is, bigger samples are more like the population from which they are drawn than small samples are.

2. RULE REPEATING CONDITION

The Law of Large Numbers is an important rule in probability and statistics theories. This rule states that as the size of a random sample increases, the sample distribution is more and more likely to get closer and closer to the population distribution. In other words, the larger the sample, the better it is as an estimate of the population.

Therefore according to this rule, when you randomly select a sample from a population the larger the sample, the better the sample is in estimating the population. Once again the larger the sample is that you draw, the better that sample is in estimating the population. This is an important rule to know in many probability and statistics problems.

3. EXPANDED EXPLANATION

Post Office Problem

When they turn 18, American males must register at a local post office. In addition to other information, the height of each male is obtained. The national average height for 18-year-old males is 69 inches.

Every day for one year, 10 men registered at a small post office and 100 men registered at a big post office. At the end of each day a clerk at each post office computed and recorded the average height of the men who registered there that day.

Which would you expect to be true? (Circle One)

(1) The number of days on which the average height was more than 71 inches was greater for the small post office than for the big post office.

(2) The number of days on which the average height was more than 71 inches was greater for the big post office than for the small post office.

(3) There is no reason to expect that the number of days on which the average height was more than 71 inches was greater for one post office than for the other.

The Post Office problem that you have just answered is one that people often have difficulty solving. On the following page is an explanation that seems to help people understand it.

LLN EXPLANATION

We often wish to find out about the characteristics of large numbers of individuals. For example, we might wish to estimate the percentage of voters in New Hampshire who are Republicans, the average family income in Suffolk County, or the average height of male students at UMass. If we were interested in the average height of the entire POPULATION of males at UMass, we could measure each one of the approximately 13,000 males on campus and find the average of these measurements. However, this would be very time consuming. We could get an estimate of the average height in the population by randomly selecting a SAMPLE from the population, and using the average of the sample as our estimate. Suppose we randomly selected a sample of 10 men and used the average height of the sample as our estimate of the average height of the population. Although we would not expect the average height of a sample of 10 men to provide a perfect estimate of the population average, it would provide some information about the population average and would give us a better estimate than if we randomly chose just one male student to base our estimate on his height.

We could see how accurate an estimate based on a sample of 10 men is by randomly selecting samples of size 10 over and over and looking at the values of the averages from each of those samples. If each sample average were a perfect estimate of the population average, then every sample average would be exactly equal to the population average. If we actually went out an obtained the averages of a large number of samples, we would find that not all of the averages would be exactly equal to the population average. Although there would be some "spread" or variability in the values of the sample averages, there would be less spread in the averages than there would be in the heights of the individual males in the population. Thus if we based our estimate on the average height of a randomly selected sample of

10 men, it is likely to more accurate than if we based our estimate on the height of one male randomly selected from the population.

Also, almost everyone has the intuition that if we wanted to get an extremely accurate estimate of the population average, we should use a large sample average as an estimate. Larger samples provide more accurate estimates of characteristics of the population than small samples do. If we selected random samples of size 100, we would find that the averages from these samples tended to be closer to the population average than the averages from samples of size 10. The averages of larger samples will tend to have values closer to the population average--that is, there will be less variability in their values because they tend to estimate the population average more accurately.

Even when an estimate is based on the average of a sample there is some chance that the estimate will be quite different than the true average of the population. For example, it is possible that the average height of a random sample of 10 men will differ from the population average by more than 2 inches. However, the bigger the sample, the less likely it is that the sample average will differ from the population average by that much.

If accuracy was our only concern, we would always base our estimates about populations on very large samples. However, when sampling is done in the real world to find out information about the population, (like in political polls), the costs of conducting the survey have to be considered as well as the accuracy. There are ways of figuring out how large a sample has to be in order to obtain a desired level of accuracy.

4. FULL TRAINING

Abstract Training

Imagine an jar that is filled with marbles. Let's say that the jar contains a very large number of marbles--thousands, millions, or larger. The marbles in this jar are known collectively as the population.

Let's say that there are two type of marbles in the jar--black marbles and white marbles. When we do this, we can now say that the population has two categories or groups, namely--black marbles and white marbles.

Now let's say that in this population of black and white marbles there are 70% white marbles and 30% black marbles. If that is the case, then we know more than that the population has two categories (black and white): we now know the proportion of the marbles in the white category (70%) and the proportion of the marbles in the black category (30%). This is known as the population distribution (other examples of distributions are 60% black and 40% white, or 85% white and 15% black, etc., but in every distribution the sum of the proportions must be 100%).

So far we know:

--A population is the entire set of objects we are interested in (all of the marbles in the jar)

--Categories refers to the types of objects in the population (black and white)

--Distribution refers to the proportion of objects in each category (70% white an 30% black in this example)

One of the major goals of statistics is to find out something about a population. More specifically, we want to find out what the population distribution is. One way that we might do this would be to actually examine all of the objects in the population and count up the number of objects in each category. In our example, we could empty the entire jar and count the number of black marbles and the number of white marbles. Using this method, we could find out exactly what the population distribution of black and white marbles was.

But, there is a very serious problem with counting all of the objects in the population: populations tend to be very large. If we were to count all of the objects in our marble population, it would take more time and effort than would be practical (imagine counting a million marbles!). So counting the entire population is impractical. What do we do instead to find out what the population distribution is?

What we do instead is to take a sample of the population. A sample is a subset of the population. We can take a sample of any size--if we pick 5 marbles, we say that the sample size is 5; if we take 60 marbles for our sample, we say that the sample size is 60, and so on.

When we take a sample from the population, we will get a distribution of the sample. The sample distribution is the proportion of objects in each category for the sample, just as the

population distribution is the proportion of objects in each category for the population. For example, if we take a sample of 10 marbles, we might get 6 blacks and 4 whites. In this case, our sample distribution would be 60% black and 40% white. We also might have happened to get 9 whites and 1 black, in which case the sample distribution would be 90% white and 10% black.

The important point here is that samples are estimates of populations. Since it is often impractical or sometimes impossible to examine the entire population, we instead have to draw samples to estimate what the population is like.

Some samples will have sample distributions that are close to the population distribution than others. For instance, in our marble example, a sample of 9 blacks and 1 white would be a very poor estimate of the population, while a sample of 8 whites and 2 blacks would be a pretty good estimate of the population.

The critical question is: What determines how likely it is that samples will give good estimates of the population? The answer is simple: if all of the samples are chosen haphazardly, or randomly (by, for example, mixing the jar and reaching into the jar blindfolded and scooping out the needed number of marbles), then there is only one factor--sample size.

This brings us to the Law of Large Numbers: as the size of a random sample increases, the sample distribution is more and more likely to get closer and closer to the population distribution. In other words, the larger the sample, the better it is as an estimate of the population.

Verbal Presentation

Now that you have all read the written explanation of the Law of Large Numbers, I thought it would be nice to demonstrate, before your eyes, that the Law of Large Numbers really does work.

So I have here (pick jar up, etc.) a genuine jar, filled with genuine black and white marbles. And as in the written explanation, there happen to be 70% white marbles and 30% black marbles in this jar.

A major purpose of statistics is to find out about a population from a sample of that population. Suppose it is your job to find out what proportion of the marbles in this jar are white and what proportion of the marbles are black. You could dump out all of the marbles and count all of them, but that would take quite a long time and wouldn't be worth the effort. For the sake of demonstration, this jar isn't very large. But if we had a very large jar filled with millions of marbles, it's easy to see how time-consuming and impractical it would be to count the entire population of marbles in the jar.

What you would probably do instead to find out what the composition of the jar was like would be to take a sample from the jar because the sample you chose would tell you something about the population; that is, the sample would be an estimate for the population.

According to the Law of Large Numbers, when you choose your sample randomly like this (reach into jar without looking, mix them up, and draw out a handful), the larger the sample, the better the sample is in estimating the population. To repeat--the

larger the sample is that you draw, the better that sample is in estimating the population.

Well , what I'm going to do now is to demonstrate the Law of Large Numbers. (reveal blackboard with summary chart) I will pick samples of size 1, 4, and 25 (gesture to the three sections of the board as you say the numbers) to show that as the sample size increases, the sample becomes a better estimate of the population.

For each sample that I draw, I will write down on the board the number of whites in the sample, the number of blacks in the sample, the percent of white in the sample, and the deviation or difference between the sample distribution and the population distribution (as you are saying the various categories, point to them on the board).

Now recall that the population distribution for this jar is 70% white, 30% black. Therefore, for example, if the sample I draw happens to be 85% white, the deviation of that sample will be 85 minus 70 or 15%, and I'll enter that number here (point to the deviation column).

I will draw a few samples of size 1, 4, and 25. After I'm done with drawing samples of each size I will calculate the average deviation of the samples from the population (point to all three "Average Deviation" boxes). The Law of Large Numbers states that as the sample size increases, the sample becomes a better estimate of the population. In other words the average deviation of the sample from the population will decrease as the sample size increases. So this number (point to "Average Deviation" box)

should go down as sample size (point to top of chart: "Sample Size= ") goes up.

So first I will draw samples of size 1. I will mix up the marbles like this (mix them while talking) and use this scooper to pick my sample (demonstrate). OK--the first marble to come out of the scoop will be my sample (Shake scoop until one drops into your hand in plain view. If more than one marble comes out, use the first one that comes out).

In this sample, I have 0 whites and 1 black (for example). (Go to the board and verbalize as you're writing down the results of the sample). So in this sample, I had 0 whites, 1 black. That means that the percent white in the sample was 0%. O minus 70% equals 70% deviation.

(go back to jar. Put the sample back, empty the scoop into the jar, and repeat the procedure. I drew 4 samples of size 1. As you go on you don't have to describe process in as much detail. After you're finished, compute the average deviation). So, the average deviation for samples of size 1 is ___%.

Now I'll pick a few samples of size 4. (Follow the same procedure as above--mix contents of jar, scoop out some marbles, pour out the first four marbles which fall into your open hand, summarize, put sample back in the jar, empty contents of scoop into jar, etc. Pick 4 samples of size 4 then compute average deviation). The average deviation for samples of size 4 is __%.

(So the same with samples of size 25. Three samples should be enough.) With this sample of 25 I have __ whites and __

blacks (write results on the board). That is __%white and the deviation is __%.

(If you draw 4 samples of size 4 and they're all 3-1, pick more samples until you have some other sample distribution. If you draw 3 samples of size 25 and you think your average deviation might be too close to the avg. deviation for samples of size 4, draw one or two more.)

Summary

The Law of Large Numbers states that as the size of your sample increases, the sample becomes a better estimate of the population. This is shown here: as the samples increased in size from 1 to 4 to 25 the average deviation of the samples from the population decreased from __% to __% to __%.

I'd like to tell you something else about the Law of Large Numbers. That is, with small samples, sometimes you can't even correctly answer the simplest questions about the population.

For example, suppose you were asked to say whether there were more white marbles in the jar or more black marbles. If you happened to draw this sample (point to a very bad sample of size 1: 0 white, 1 black. If there isn't one, then go to the samples of size 4 and point to a 1 white, 3 black or 2 white, 2 black) you would say "Well, from my sample, I think there are more blacks than whites and I would be wrong. But look at the larger sample of size 25: you can always correctly answer at least the most basic

question-- 'Are there more whites or more blacks?' With smaller samples, that is not always possible.

So I've demonstrated the Law of Large Numbers--as the sample increases in size, the sample becomes a better estimate of the population.

Explained Examples

At a large urban university, a student organization used the computer to select 1500 student numbers at random, and contacted those selected to ask whether they were willing to have $20 added to their fees for the construction of a new gym. About 72% of the 1500 students questioned said they would favor such a fee in order to have expanded athletic facilities. Although the Dean of Student Services at the university agreed the $20 per student would probably be adequate to pay the annual mortgage on a new gym, he argued that: "Since there are a very large number of students at our university who are from lower and lower-middle class families, a $20 fee would be quite a hardship on them. Thus, it is very unlikely that a majority of students at the university would be willing to have $20 added to their fees for the new gym. While a majority of the people you asked were in favor of the fee, it is far from certain that a majority of the entire student population is in favor of the fee. A great many other people have not been consulted."

Do you agree with the Dean that it is "far from certain" that a majority of the student population favors the fee for the new gym? Explain.

Please consider this problem for a few moments. After you have considered the problem and analyzed it for a minute or two, turn the page for an analysis.

The student organization is trying to assess the attitude of the students at the university toward spending $20 per student to build the new gym. In terms of the Law of Large Numbers, they are trying to find out the population distribution of the university students' attitudes toward the fee increase. To do this, the organization randomly selected 1500 students--a sample of size 1500--and asked them if they were in favor of the fee for the new gym. Of the 1500, 72% were in favor of it (and 28% were against it). According to the Law of Large Numbers, which states that the larger the sample, the better it is in estimating the population, there is overwhelming evidence that a majority of the student population favors the fee. Recall that in the gumball demonstration, samples of size 25 were very good estimates of the population. Thus, it can be concluded that a majority of students at the university are if favor of the fee.

What about the Dean's argument that it is unlikely that a majority of the student population favors the fee because they would consider a $20 fee a hardship? Although this argument may have intuitive appeal, it should be discounted because it is not supported by any data, and is in fact contradicted by the large sample of 1500 students.

A major New York law firm had a history of hiring only graduates of large, prestigious law schools. One of the senior partners decided to try hiring some graduates of smaller, less prestigious law schools. Two such people were hired. Their grades and general record were similar to those of people from the prestigious schools hired by the firm. Although their manners and "style" were not as polished and sophisticated as those of the predominantly Ivy League junior members of the firm, their objective performance was excellent. At the end of three years, both of them were well above average in the number of cases won and in the volume of law business handled. The senior partner who had hired them argued to colleagues in the firm that, "This experience indicates that graduates of less prestigious schools are at least as ambitious and talented as graduates of the major law schools. The chief difference between the two types of graduates is in their social class background, not in their legal ability, which is what counts."

Comment on the thinking that went into this senior partner's conclusion. Is the argument basically sound? Does it have weaknesses? (Disregard your own initial opinion, if you had one, about graduates of non-prestigious law schools, and concentrate on the thinking that the senior partner used.)

Please consider this problem for a few moments. After you have considered the problem and analyzed it for a minute or two, turn the page for an analysis.

The senior partner is trying to draw a conclusion about a certain population. We can think of the members of this population as newly graduated lawyers, from nonprestigious law schools, who otherwise meet the law firm's hiring standards. If we divide the members of this population into two categories, "excellent" and "mediocre or worse," we can think of the population distribution as the % in each category. The senior partner has concluded that the % in the "excellent" category is very high, or anyway, just as high as in another population, involving graduates of prestigious law schools. This conclusion was based on observing a sample of size 2, in which the sample distribution was 100% "excellent," 0% "mediocre or worse."

Apart from any other considerations, however, the sample distribution for size 2 is apt to be quite different from the population distribution: the latter could be only 60% or 50% or even perhaps as low as 40% "excellent," and a 2-0 sample split would not be so unusual; just as one would not be at all amazed to draw 2 out of 2 red gumballs from an urn with only 40% reds. So the senior partner's attitude is quite unwarranted: a larger sample is needed.

APPENDIX B

TESTING MATERIALS

Fong et al. (1986) open ended problems

Slot Machine Problem

For his vacation, Keith decided to drive from his home in Michigan to California to visit some of his relatives and friends. Shortly after crossing the border to Nevada, Keith pulled into a gas station and went inside to buy a state map. There, in the corner of the gas station, were two slot machines. Keith had heard about slot machines,before, but had never actually seen one. He went over to the slot machines and looked at them, trying to figure out how they worked. An old man who was sitting close to the machines spoke to Keith. "There ain't no winning system for slot machines. It's all luck. You just put in a coin, pull the lever, and hope that you'll win. But let me tell you this: some machines are easier to lose on than others. That's because the owners can change the mechanism of the slots so that some of them will be more likely to make you lose. See those two slot machines there? The one on the left gives you about an even chance of winning, but the one on the right is fixed so that you;ll lose much more often than you win. Take it from me-I've played them for years." The old man then got up and walked out of the gas station.

Keith was by now very intrigued by the two slot machines, so he played the machine on the left for a couple of minutes. He lost almost twice as often as he won. "Humph," Keith said to himself. "The man said that there was an even chance of winning at the machine on the left. He's obviously wrong." Keith then tried the machine on the right for a couple of minutes and ended up winning more often than be lost. Keith concluded that the man was wrong about the chances of winning on the two slot machines. He concluded that the opposite was true-that the slot machine on the right was more favorable to the player than the machine on the left.

Comment on Keith's conclusion and his reasoning. Do you agree? Explain your answer.

IRS Problem

Martha was excited about her new job with the IRS. She knew that she had been well trained but dreaded spotchecking incoming tax returns to see what kinds of errors the taxpayers were making. Every serious error she found meant extra paperwork for her and she already felt overworked. "I suppose things just get worse as the deadline approaches," she mentioned to her friend Laura, who had been on the job for several years. "I guess that people who file near April 15 tend to be in a terrible hurry and make all kinds of mistakes." "That certainly is my impression," said Laura. "I must have processed hundreds of returns last year and found that almost one-quarter of them had serious errors. Mind you, maybe things will be better this year. There's been a lot of publicity about tax reform and quite a few TV spots reminding people to get started working on their tax returns early."

When they met for lunch a few weeks later, Martha seemed quite relieved. "Something must be working," she said to Laura. "Even though there are just about as many returns filed at the last minute as there were last year, when I picked out ten of them and went through them carefully, I only found one serious mistake. It doesn't look like things are going to be nearly as bad as I had feared."

Comment on Martha's conclusion and her reasoning. Do you agree? Explain your answer.

Multiple choice center version problems

Book Problem

An investigator studying some properties of language selected a paperback and computed the average word-length in every page of the book (i.e., the number of letters on that page divided by the number of words). Another investigator took the first line in each page and computed the line's average word-length (i.e., the number of letters on the first line divided by the number of words). The average word-length in the entire book is 4 letters. However, not every line or page has exactly that average. Some may have a higher average word-length, some lower.

The first investigator counted the number of pages that had an average word length of between 3 and 5 and the second investigator counted the number of first lines on a page that had an average word length of between 3 and 5.

Which of the following would you expect to be true?

(1) The page investigator had a higher percentage of word-length averages between 3 and 5.

(2) The line investigator had a higher percentage of word-length averages between 3 and 5.

(3) There is no reason to expect that the percentage of word-length averages between 3 and 5 would be larger for one investigator than for the other.

Please write down your reasons for selecting the answer that you did.

Blood Type Problem

When a company is hiring new employees often they collect medical information about the new employee. Part of the information companies gather is the blood type of each new employee (in case of an emergency in the workplace). It is known that the incidence of type O blood in the population of the U.S. is around 45%, and the remaining 55% have either type A, B, or AB.

A small company, Smith's Steamroller Inc., hires 10 new people a month. A much larger company, Jones Farm Equipment Inc., hires 50 new employees per month. Both companies have recorded the percentage of type O new employees per month for the last several years. Which would you expect to be true? (Circle one)

(1) The number of months in which between 35% and 55% of the new employees has blood type O was greater for Smith Steamroller than for Jones Farm Equipment.

(2) The number of months in which between 35% and 55% of the new employees had blood type O was greater for Jones Farm Equipment than for Smith Steamroller.

(3) There is no reason to expect the number of months in which between 35% and 55% of the new employees had blood type O to be greater for one company than the other.

Please write down your reasons for selecting the answer that you did.

Tennis Problem

Two tennis players, Stephanie and Matt were arguing about how many people they thought played tennis in the U.S. Stephanie said that she thought very few Americans played because there are not many big United States tournaments, and schools tend to have very small tennis teams. Matt insisted that because almost everyone he knew played tennis, at least half of all Americans must play tennis.

Stephanie went to the library and found the Statistical Abstract of the U.S. 1989. In this publication the Bureau of the Census includes statistics about the activities Americans enjoy. It states the 15% of Americans play tennis.

After reading this report Matt thought that it was ridiculous. He and Stephanie decided to take a random poll on the streets of Boston in order to find out how many Bostonians play tennis. Every day for 2 weeks they each conducted their own survey. Each day Matt stopped 10 people and Stephanie stopped 100.

Which of the following would you expect to be true?

(1) Stephanie had more days in which she found between 10% and 20% of the people surveyed played tennis than Matt did.

(2) Matt had more days in which he found between 10% and 20% of the people surveyed played tennis than Stephanie did.

(3) There is no reason to expect that either one of them had more days in which they found between 10% and 20% of the people surveyed played tennis.

Please give your reasons for selecting the answer that you did

Hospital Problem

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it maybe higher than 50%, sometimes lower.

For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys.

Which of the following would you expect to be true?

(1) The number of days on which more than 60% of the babies born were boys was greater for the smaller hospital than for the larger hospital.

(2) The number of days on which more than 60% of the babies born were boys was greater for the larger hospital than for the smaller hospital.

(3) There is no reason to expect that the number of days on which more than 60% of the babies born were boys should be greater in one hospital than in the other.

Please write down your reasons for selecting the answer that you did.

Tail version problems

Geology Problem

In a particular geology course, students must frequently determine the weight of rock samples. To develop the needed skills, one instructor has her students practice weighing an object many times on the same scale. The object, a metal disk is known to weigh exactly 1000 grams. Although the scale is not completely accurate, it is equally likely to read high as it is to read low. However, it is never off from the actual weight by more than 25 grams.

After a lot of practice, each student in one of her classes (section 1) weighed the disk 20 times and then computed the average of the 20 weighings. In the other class (section 2), after a lot of practice, each student weighed the disk 5 times and then computed the average of the 5 weighings.

Which would you expect to be true? (Circle One)

(1) The percentage of students who obtained an average weight of more than 1010 grams was larger in section 1 than in section 2.

(2) The percentage of students who obtained an average weight of more than 1010 grams was larger in section 2 than in section 1.

(3) There is no reason to expect that the percentage of students who obtained an average weight of more than 1010 grams was larger for one section than the other.

Please write down the reasons that you selected the answer you did.

Women's Heights Problem

At a small women's liberal arts college the students are required to have a physical examination each year during the first week of the school year. Physicals are offered at both the main health services center and at a small clinic that is part of the recreation center. Each day during that first week of the semester 50 women go the main health clinic for the physicals and 10 women go to the smaller clinic.

Part of the physical examination includes taking the women's heights and weights. The national average height for college age women is 5'4". At the end of every day a nurse at each clinic records the heights of the women that visited the health clinic that day and calculates the average.

Which of the following would you expect to be true?

(1) The number of days on which the average height was less than 5'1" is greater for the small clinic than for the large clinic.

(2) The number of days on which the average height was less than 5'1" is greater for the large clinic than for the small clinic.

(3) There is no reason to expect the number of days on which the average height was less than 5'1" to be greater for one clinic than for the other.

Please write down the reasons that you selected the answer you did.

Lures

Pollution Problem

It is well known in the auto industry that the amount of pollution produced by new cars varies slightly from car to car because of slight differences in the machining of parts and in assembling. A certain amount of variation is accepted and the plant officials do not worry unless a sizable number of cars exceed a maximum pollution threshold.

A certain plant maintains two assembly lines, on that produces 25 cars a day and one that produces 100. Both lines use the same parts and the workmen on both lines have equivalent training.

Last week, each car produced in the plant was tested. It was found that about 8% (about 2 per day) of the cars produced on the small line exceeded the pollution threshold. Without any further information, what would you expect the percentage of cars on the large assembly line that exceeded the pollution standards to be?

(1) Greater than 8% .

(2) Less than 8%.

(3) I have no reason to think it should be greater than or less than 8%.

Please write down your reasons for selecting the answer that you did.

Treasury Department Problem

The Treasury Department monitors people's incomes by taking random samples from a large set of data on incomes provided by the Internal Revenue Service. Tom took one random sample of 25 incomes from these data and George took a random sample of 200 incomes from the same set of income data. Each calculated the percent of people in his sample who earned over $50,000.

Which would be true?

(1) The percent of people who earn over $50,000 in the sample of 200 is likely to be greater than the percent of people who earn over $50,000 in the sample of 25.

(2) The percent of people who earn over $50,000 in the sample of 25 is likely to be greater than the percent of people who earn over $50,000 in the sample of 200.

(3) There is no reason to expect that the percent of people who earn over $50,000 will be greater in one sample than in the other.

Please write down your reasons for selecting the answer you did.

Fish Hatchery Problem

The manager of a fish hatchery monitors data about the length and the weight of trout that are raised in the hatchery tanks. This information is important because it has been found that if the trout are too small when they are released into rivers and lakes, the survival rate will be low. Since There are thousands of trout in the tanks, the data are obtained by taking random samples of the trout in each tank. From past data, they know the average weight of trout at a certain age is one pound. On a given day, two employees each take a random sample of trout at that age to measure. However, one takes a sample of 10 and the other takes a sample of 50. They weigh each trout in the sample and compute the percent of trout that are less than 3/4 of a pound.

Which would be true?

(1) The percent of trout less than 3/4 of a pound should be greater in the small sample than in the large sample.

(2) The percent of trout less than 3/4 of a pound should be greater in the large sample than in the small sample.

(3) There is no reason to expect that the percent of trout less than 3/4 of a pound would be greater in one sample than in the other.

Please write down your reasons for selecting the answer that you did.

Personality Scale Problem

One of the instruments used for answering personality development is the Inventory of Psychosocial Growth (IPG). The IPG consists of 80 items, each of which is thought to correspond to one aspect of maturity. Overall scores on the IPG can range from 0 to 80 and it has been found over the years that the average score for college students is 53. Last year at UMass the IPG was administered to two samples of randomly selected undergraduates. One sample included 30 students, the other included 200 students.

It was found that about 13% of the students in the small sample received scores greater than 70. Without any further information, what would you expect the percentage of students from the larger sample that scored over 70 to be?

(1) Less than 13%

(2) Greater than 13%

(3) I have no reason to think that it should be less than or greater than 13%.

Please write down your reasons for selecting the answer that you did.

Brewery Problem

A brewery buys nearly all of its reusable glass bottles from a local glass manufacturer. One summer, however, the local company is unable to deliver enough bottles, and the brewery orders a shipment from a large glass manufacturer that distributes its products nationwide. On the first day that these new bottles are used, however, the bottle-filling machinery has to be stopped four times because of jamming, and as a result, production for the day is unusually low. (Ordinarily the brewery does not experience more than one jamming stoppage per day and frequently there are none at all.) The foreman is worried about the new bottles. He decides to test the new bottles produced by the national manufacturer carefully. He randomly selects 300 cases of these new bottles and instructs the bottle-filler operators to record carefully each jamming incident. Meanwhile, company mechanics carefully lubricate and check adjustments on the bottle-filling machinery. When they are finished, the bottle-filling machinery is running more smoothly than it has in years. During the next 2 days, the 300 cases of new bottles are fed to the machine. There are only two jamming incidents, one each day. The foreman concludes that there is in fact little or no real disadvantage of the new bottles with respect to jamming of the bottle-filling machine.

What do you think about the foreman's reasoning? Give reasons for your answers.

Auditor Problem

An auditor for the Internal Revenue Service wants to study the nature of arithmetic errors on income tax returns. She selects 4000 Social Security numbers by using random digits generated by an "Electronic Mastermind" calculator. And for each selected social security number she checks the 1988 Federal Income Tax return thoroughly for arithmetic errors. She finds errors on a large percentage of the tax returns, often 2 to 6 errors on a single tax return. Tabulating the effect of each error separately, she finds that there are virtually the same number of errors in favor of the taxpayer as in favor of the government. Her boss objects vigorously to her assertions, saying that it is fairly obvious that people will notice and correct errors in favor of the government, but will overlook errors in their own favor. Even if her figures are correct, he says, looking at a lot more returns will bear out his point.

Comment on the auditor's reasoning and her boss's contrary stand. Which do you think is correct? Please give reasons for your answers.

APPENDIX C

THREE POINT CODING SYSTEM

Due to the fact that the Fong et al. (1986) materials we were given describe their coding system as consisting of a four point scale, rather than the three point scale they report in their journal article, and the fact that the problems for the Fong and Nisbett (1991) study were somewhat different from the problems in the present study, the three point coding in the present study attempted to replicate what I thought would be as close as possible to how Fong and Nisbett would have coded the answers to the problems in the present study, based on what they have previously reported as their coding scheme. A description of the coding scheme used in the present study is listed below.

(1) A one was given for an answer which was entirely deterministic and contained nothing probabilistic whatsoever. Answers in which it was extremely difficult to understand what the subject meant were also given this score.

(2) A two was given for an answer in which there was some mention of a probabilistic principle, but the intention in mentioning it was unclear. Also answers that were a mix of probabilistic and deterministic reasoning, but seemed to rely on the deterministic were given a two. Answers in which the subject inappropriately used a probabilistic rationale were also given a score of two.

(3) A score of three was given to answers in which a clear, or fairly clear probabilistic principle was appropriately used in an answer to the problem.

APPENDIX D

RATIONALES CODING SYSTEM

I. No regard for sample size

1 Any sample is as good as any other

size doesn't matter, samples are good, either samples is okay, they are both samples, etc.

a) because they are both from the same population.

2 As long as samples are random, nothing else about the sample matters.

Like sample size, etc.

3 If you are talking about percents or averages, not numbers, sample size doesn't matter.

a)because they are both from the same population.

II. Bigger is somehow better

4 Bigger samples give bigger percents.

Because it is a bigger sample you will get a bigger percent.

5 Bigger samples give greater chances for getting anything.

Because it is a bigger sample you have a greater chance of getting...

6 Because it's bigger

Just this phrase, because ____ is bigger than...

III. No regard for data

7 The given data is not important here.

An explanation that mentions factors other than the data or numbers given (a non statistical answer) as being the causal reason. In FA, no sample mentioned.

8 It's just luck

The outcome is due to luck, chance, coincidence. It's random, who knows?

9 You can't predict anything,

a) because they are from the same population, or the same factors are at work, etc.

b) because variability is involved.

c) because there is not enough information given.

10 You can't make predictions unless you know about the whole population.

IV. Small samples are inadequate

11 You can't conclude much from a small sample.

(An explanation that really doesn't go beyond this statement)

12 With small samples you are more likely to get inaccurate results.

Small samples can lead to extreme, strange, inaccurate, etc. results

13 With a small sample extreme values may occur w/o other values to balance them out.

This explanation is a little bit beyond #12, it indicates some understanding of why #12 is true.

V. Big samples are ...

14 Big samples are representative of the population.

Simply this sort of statement, no mention of superiority of large samples, often applies in FA or lures. i.e. Because of the sample, I believe, etc.

15 With large samples extreme values tend to be balanced out, thus there is less chance for extreme avg.s, etc.

Understanding beyond 18, some explanation for why 18 and/or 19 are true. The other end of number 13.

VI. Some comparison between samples

16 Big samples are better than small samples.

This very simple statement, the kind you want to be able to read into, but it doesn't say enough to be able to really.

17 Big samples are more accurate or representative than small samples.

a) Frequently just a regurgitation of something they have been told or have read, doesn't really say what it means to be more accurate, however it does say more than 17.

b) Something more than "bigger is more accurate." Some explanation as to what it means to be more accurate, i.e., "the bigger one is more like the population, more representative of the true values, etc."

18 You can make better conclusions and predictions about a pop. from a big sample than from a small one.

Indicates some understanding of what one can do with a more accurate sample.

19 The %s or numbers for the samples may be different, but no way to know which way.

Frequent normative response for lures or FAs, indication that there is not enough information available to predict which sample is going to have bigger or smaller percentage, etc.

20 Big samples have less deviation/variability from the average, etc., than small samples.

Also vice versa: Small samples deviate more from the average., etc.

VII. Other

21 an idiosyncratic answer:

something that can somehow been understood, whether normative or incorrect, but is an atypical answer.

22 unclassifiable

an answer that really says nothing. It makes no sense or is merely a repetition of the question, etc.

BIBLIOGRAPHY

Cheng, P.W. & Holyoak, K.J. Pragmatic Reasoning Schemas *Cognitive Psychology*, 1985, 17, 391-416.

Cheng, P.W., Holyoak, K. J., Nisbett, R.E., & Oliver, L.M. Pragmatic versus Syntactic Approaches to Training Deductive Reasoning *Cognitive Psychology*, 1986,18, 293-328.

Evans, J. St. B.T. *The psychology of deductive reasoning*. London: Routledge & Kegan Paul, 1982.

Fong, G.T., Krantz, D. H., & Nisbett, R.E. The Effects of Statistical Training on Thinking about Everyday Problems. *Cognitive Psychology*, 1986,18, 253-292.

Fong, G.T. & Nisbett, R. E. Immediate and delayed transfer of training effects in statistical training. *Journal of Experimental Psychology: General*, 1991, 120, 34-45.

Gick, M.L., & Holyoak, K.J. Analogical Problem Solving. *Cognitive Psychology*, 1980,12, 306-355.

Gick, M.L., & Holyoak, K.J. Schema Induction and analogical transfer. *Cognitive Psychology*, 1983, 15, 1-38.

Griggs, R.A., & Cox, J.R. The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, 1982, 73, 407-420.

Holland, J. H., Holyoak, K.J., Nisbett, R.E., & Thagard, P.R. *Induction. Processes of Inference, Learning & Discovery*. M.I.T. Press, 1986.

Holyoak, K.J., Koh, K., & Nisbett, R.E. A theory of conditioning: Inductive learning within rule-based default hierarchies. *Psychological Review*, 1989, 96 (2), 315-340.

Johnson-Laird, P. N., Legrenzi, P., & Sonino-Legrenzi, M. Reasoning and a sense of reality. *British Journal of Psychology*, 1972, 63, 395-400.

Kahneman, D. & Tversky, A. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 1972, 3, 430-454.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge Univ. Press, 1982.

Katona, G. *Organizing and memorizing.* New York: Columbia University Press, 1940.

Krantz, D.H., Fong, G.T., & Nisbett, R.E. *Formal training improves the application of statistical heuristics to everyday problems.* (unpublished).

Kunda, Z. & Nisbett, R.E. Prediction and the partial understanding of the law of large numbers. *Journal of Experimental Social Psychology*, 1986, 22, 339-354.

Nisbett, R.E., Fong, G.T., Lehman, D.R., & Cheng, P.W. Teaching Reasoning. *Science*, 1987, 238, 625-631.

Nisbett, R.E., Krantz, D.H., Jepson, C., & Kunda, Z. The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 1983, 90, 339-363.

Nisbett, R.E., Krantz, D.H., Jepson, C., Fong, G.T. *Improving inductive inference* in Kahneman et al. (1982).

Ploger, D. & Wilson, M. Statistical Reasoning: What is the role of inferential rule training? Comment on Fong and Nisbett. *Journal of Experimental Psychology: General*, 1991, 120, 213-214.

Reich, S. S. & Ruth, P. Wason's selection task: Verification, falsification and matching. *British Journal of Psychology*, 1982, 73, 395-405.

Singley, M.K., & Anderson, J.R., *The Transfer of Cognitive Skill.* Cambridge, Mass.: Harvard University Press, 1989.

Thorndike, E.L. & Woodworth, R.E. The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review,* 1901, 8, 247-261.

Thorndike, E.L. *Principles of teaching.* New York: A.G. Seiler, 1906.

Tversky, A., & Kahneman, D. The belief in the "law of small numbers." *Psychological Bulletin,* 1971, 76, 105-110.

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science,* 1974, 185, 1124-1131.

Wason, P.C. Reasoning. In B.M. Foss (Ed.) *New horizons in psychology 1.* Harmondsworth, England: Penguin, 1966.

Wason, P.C., & Johnson-Laird, P.N. *Psychology of Reasoning: structure and content.* Cambridge, Mass.: Harvard University Press, 1972.

Well, A., Pollatsek, A., & Boyce, S. Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes.*, 1990, 47, 289-312.